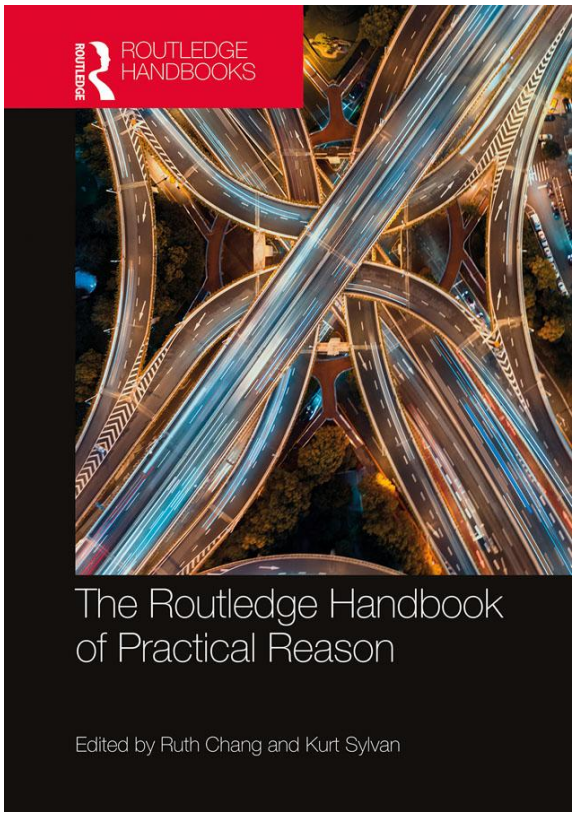
The Routledge Handbook Of Practical Reason

Routledge Handbooks in Philosophy

Kurt Sylvan & Ruth Chang



2020

Contents

The Routledge Handbook of Practical Reason 20
Routledge Handbooks in Philosophy 20
Title Page 21
Copyright 23
Dedication 24

Acknowledgments **25**

Contributors **27**

An introduction to the philosophy of practical reason **30**

I What is the philosophy of practical reason? **33**
1 The philosophy of practical reason as the philosophy of action, mind, or moral psychology 33
2 The philosophy of practical reason as metaethics and ethics 35
3 The philosophy of practical reason as theory of rationality 41

II A guide to the volume **43**
1 Foundational matters 43
2 Practical reason in the history of philosophy 44
3 Practical reason, action theory, and moral psychology 45
4 Practical reason and normativity 47
a The nature of normative practical reasons 48
b High-level substantive matters 51
5 Practical rationality 52

References **56**

Part 1: Foundational matters **61**

1. Some central questions about practical reason 62

1 Introduction 64

2 Common objections to normative realism 65
2.1 Metaphysics 65
2.2 Motivation 66
2.3 Epistemology 67

3 Reasons and rationality 70
3.1 Three ideas of rationality 70
3.2 The normativity of requirements of (structural) rationality 72
3.3 Anchoring reasons in an agent's attitudes 73

4 The relation between normative truths and non-normative truths 77

5 A very brief reiteration 79

References 80

2. Practical reason 82

1 Introduction 84

2 Rationality 85

3 Normativity 89

4 Rationality is not the same as normativity 92

5 Responses 95
Does what you ought supervene on the mind? 95
Does normativity supervene on the mind? No 97
Does rationality supervene on the mind? – Yes 98

6 A Kantian response 100

7 Summary 103

Note 104

References 105

3. Can reason be practical? Narrow and broad conceptions and capacities **107**

Introduction **109**

1 Narrow and broad conceptions of practical reason **110**

2 From a narrow to a broad conception – Kant **112**

3 Extending the broad conception – Aristotle **118**

3 Directions of fit – Hume **121**

4 Psychological realism? **128**

5 Conclusion **130**

References **131**

4. Practical reason and social practices **133**

1 Introduction **135**

2 Methodology, motivation, and non-ideal theory **137**

3 Culture: giving shape and content to our thinking **141**

4 Social practices **144**

5 How, again, is this relevant to the study of practical reason? **149**

Acknowledgments **152**

References **153**

5. How to be a pragmatist **156**

1 Pragmatism and the limits of normative intuitions **158**

2 Pragmatist methods as problem solving **163**

3 An example: recognition of slavery as a moral wrong **167**

4 Updating the pragmatist research program **173**

6. What is it to be a rational agent? 175

1 A grounding framework for reasons **179**

2 The sources of normative reasons **181**

3 Source voluntarism about normative reasons **183**

4 What do we will when we create reasons? **185**

5 Some upshots for normative reasons and rational agency **188**

6 The problem of well-formed choice situations **191**

7 The activist view in action **196**

References **197**

Part 2: Practical reason in the history of philosophy 199

7. Practical reasoning in early Chinese philosophy 200

Contextual reasoning in Confucianism **202**

Reasoning from concrete experience **204**

Analogical reasoning in the *Mencius* **206**

How reflection, reasoning, feeling and desire interweave and mutually influence one another in the *Mencius* **210**

Mencius on exercising judgment when values conflict **212**

Practical reasoning and Zhuangzi's intuitive activity **214**

Synthesizing Confucian and Daoist practical reasoning **219**

Further reading **220**

References **221**

8. Aristotle on deliberation 223

I A geometrical model for deliberation **226**

Backwards reasoning **233**

Why can't the vicious person reason well? **236**

(P1) Vicious people have, as their ends, things that are not good for them **237**

(P2) Good deliberation preserves the status of your end as something good for yourself **239**

C: The vicious person cannot reason well, because you can't preserve what is not there **240**

Deliberating immersively, not reflectively **242**

Works cited **246**

9. Hume's robust theory of practical reason 248

Introduction **250**

Elements of a theory of practical reason **251**

Hume on the *standards* for practical deliberation **253**

Hume on the *capacities* required for the activity of practical deliberation **260**

Hume on the activity of practical deliberation **267**

In sum **273**

References **274**

10. Kant's approach to the theory of human agency **277**

I Introduction **279**

II Two dialogues **281**
Leibniz and Clarke 281
Bratman and Korsgaard 282

III A question of method **285**

IV Kant's critical method **287**

V Kant's theory of agency **290**

VI Why Kant thinks his method tells us what agency is **292**

VII Conclusion **295**

Further reading **296**

References **297**

11. Anscombe on acting for reasons **298**

Introduction **300**

1 Intentional action and acting for reasons **302**

2 Reasons and causes **305**

3 Reasons as constituents of action **309**

4 Anscombe's view in context **313**

References **317**

Part 3: The philosophy of practical reason as action theory and moral psychology **319**

12. Three dogmas of agency theory **320**

Deliberation **323**

Endorsement **326**

Identification **329**

Conclusions **331**

References **334**

13. Some reflections on the relationship between reason and the will **335**

The conditions of action as limits on the possibility of practical irrationality **338**

Further thoughts about the metaphysical limits on irrationality **344**

The possibility of passive agency **347**
Two case studies 347
Other cases of passive agency 351

The desirability and necessity of passive agency **357**

14. Three for the price of two **360**

1 The first view: holism in the theory of reasons **362**

2 The second view: realism in the theory of motivating reasons **364**

3 The favouring relation **365**

4 Aristotle on practical reason **366**

5 The third view **368**

6 An example of theoretical reasoning **370**

7 Explaining reasons and reasoning **372**

8 Formally valid reasoning **373**

9 Intermediate summary **375**

10 Getting it wrong **376**

11 First difficulty **377**

12 Second difficulty **379**

13 Third difficulty **380**

14 A final challenge **382**

References **383**

15. The guise of the good 384

Background and motivation **386**

The nature of the guise **390**

Common objections and replies **393**
Purported counterexamples: Akrasia 393
Purported counterexamples: perversion 395
Theoretical difficulties: alternative constitutive aims 396
Theoretical difficulties: superfluity 397

Conclusion **398**

References **399**

16. Motivational internalism and externalism 402

Motivational internalism **405**

Conditional vs unconditional internalism **407**

Direct vs deferred internalism **410**

De dicto vs de re internalism **412**

Motivational internalism as an empirical thesis **414**

Unconditional internalism revisited **418**

Conclusion **421**

References **422**

17. Emotions in practical reasoning 425

1 Affect as evaluative **427**

2 Warrant for emotion **429**

3 Reconstructing emotional reasoning **432**

4 Discounting reasons **435**

5 Emotions and higher order reasons **438**

5 Conclusion **441**

References **442**

18. Psychopathy, agency, and practical reason 444

1 Psychopathy **446**

2 Psychopathy and practical reason in philosophical discourse **449**
2.1 Moral rationalism and moral sentimentalism 449
2.2 Motivational judgment internalism and motivational judgment externalism 452
2.3 Morally responsible agency: rational competence versus moral competence 454

3 Progress and future directions **459**

References **462**

19. Practical reason and social science research 466

1 Rationalism and sentimentalism about moral judgments **468**
Emotions influence and cause moral judgments 468

Emotions are necessary for moral judgment 470
Reasoning does not play the role that the rationalist thinks it does 472

2 Reasons, action, and agency **475**

3 Good reasoning and overcoming bias **479**
Hidden mechanisms 480
Broader lessons about practical reasoning from the epistemology of implicit cognition 481

4 Conclusion **483**

References **484**

Part 4: The philosophy of practical reason as the theory of practical normativity **488**

Section 1: The nature and grounds of normative practical reasons **489**

20. Objectivism about reasons **490**

1 Two kinds of theory of practical reasons **493**

2 Subjective theories of reasons **496**

3 Why subjective theories are mistaken: the agony argument **499**

21. How to be a subjectivist **506**

What is valuing? **508**

Partners in crime? **511**

Why subjectivism? **512**

The thin end of the wedge for the subjectivist **515**

Agony, blades of grass, and all that **517**

Reasons to be moral **520**

Conclusion **523**

22. Kantian constructivism 524

1 Introduction **526**

2 Motivations for normative constructivism **527**

3 Varieties of normative constructivism **529**

4 Constructing universalizability requirements **532**

5 Constructing respect for persons **534**

6 Constitutivist supplements **539**

7 Reasoning and other people **542**

23. Constitutivism 550

1 Introduction **552**

2 Smith's constitutivism **555**

3 Shmagency's revenge **558**

4 Is it just the open question argument all over again? **563**

5 Equivocations **565**

6 Methodology **568**

7 Conclusion **570**

References **571**

24. Reasoning first 573

Current thinking **575**

The first difficulty **577**

The second difficulty **580**

The third difficulty **584**

Alternative **586**

Identifying the wrong kind of reasons and the "voluntary" **588**

Explaining actions done for reasons **591**

The mysterious "normative" relation **593**

Conclusion **596**

References **597**

25. Normative nonnaturalism 600

1 Meta-ethics **604**

2 Ontology **608**

3 Concepts and properties **612**

4 Against normative naturalism **616**
a The co-extensiveness argument 617
b The normativity objection 618
c Scientific analogies 620

5 Railton's agreement **630**

6 Gibbard's agreement **631**

Section 2: Some substantive matters 632

26. Non-requiring reasons 633

1 Introduction **635**

2 The standard view **637**

3 Permissibility-conferring reasons **639**

4 Commendatory reasons **643**

5 Conclusion **646**

References **650**

27. Requirements of reason 652

1 Requirements of reason: the basic idea **654**

2 Explaining requirements of reason **658**

3 Explaining moral requirements **662**

4 Toward a relational account **665**

5 Conclusion **668**

28. Normative pluralism and skepticism about 'ought *simpliciter*' 670

1 Distinguishing the pluralist view from the unified view **673**

2 Can there be conflict situations? **677**

3 An argument for the pluralist view **680**

4 Objections **684**

5 Oughts *simpliciter* and the strong pluralist view **686**

6 "Metaphysical" and "Evaluative" Propositions about *Ought* **693**

7 Objections to the strong pluralist view **696**

8 Conclusion **700**

Acknowledgments **701**

References **702**

29. There is no moral *ought* and no prudential *ought* 704

1 Introduction **706**

2 The common view: there is a moral *ought* and a prudential *ought* **708**

3 Moral *ought* statements do not just state moral requirements **710**

4 Sometimes (often), morality does not recommend that one do the morally best thing one could do **713**

5 Sometimes morality's recommendation is to do a morally good thing that is not the agent's morally best option **716**

6 Objection: the moral perspective ignores self-regarding reasons **719**

7 Objection: something can be good advice although only a jerk would offer it **721**

8 Making sense of morally, you ought to do it 723

9 Prudence does not always recommend the option that is best for the agent **726**

10 Making sense of prudentially, you ought to do it 728

11 Objections from other *oughts* **729**

12 Summary **731**

13 Why this matters **733**

30. Practical reason and the second-person standpoint 736

1 Introduction **738**

2 Second-personal reasons **742**

3 Deontic second-personal reasons **744**

4 Morality's authority **749**

5 Conclusion **751**

References **752**

Part 5: The philosophy of practical reason as the theory of practical rationality **754**

31. The normativity of rationality **755**

1 Coherence and rationality **757**

2 Normative force and normative reasons **760**

3 The normative impotence of coherence requirements **762**
3.1 The normative impotence of narrow-scope coherence requirements 762

3.2 The symmetry argument and its discontents **764**
3.2.1 The argument 764
3.2.2 Criticisms of the symmetry argument 765
3.3 The missing reasons argument 766

4 Responding to the challenge **769**
4.1 Alternative views of normative force 769
4.2 Alternative views of rationality 771

5 Conclusion **773**

References **774**

32. The eclipse of instrumental rationality **777**

Introduction **779**

1 Instrumental rationality: some preliminaries and disclaimers **781**
1.1. The face-value understanding of instrumental rationality and reasoning . 781
1.2 Some contrasting phenomena 782
1.3 Face-value instrumental reasoning and rationality in more detail 784
1.4 Claims and disclaimers 785

2 Against the Status Explanation Claim **787**
2.1 Argument from the explanation of canonical examples 787
2.2 Argument from alienation 788
2.3 Technical knowledge doesn't mark out a distinct field of practical rationality 790
2.4 Arguments from subsumption and embeddedness 792
2.5 Argument from the value of rationality 795
3 Against the Necessary Glue Claim 797
4 The non-instrumental structure of practical reason 799
5 The non-instrumental unity of reason 803

Bibliography **807**

33. Rationality, regret, and choice over time 811

Introduction **813**

Going wrong **814**

Temptation and temporary preference reversals **815**

Vague goals and procrastination **818**

Incommensurable alternatives and gratuitous costs **822**

Conclusion **825**

References **826**

34. Plan rationality 828

Plan rationality? **830**

The strategy of self-governance **834**

Extending the strategy of self-governance: 'acting together with oneself' over time **836**

Shuffling **838**

Temptation **839**

A unified account **841**

A reason for, and end of, self-governance? 842

The significance of a stable, rational equilibrium? 844

Note 845

References 846

35. Between sophistication and resolution – wise choice 849

1 Violation of Independence and dynamic inconsistency 853

2 Sophisticated choice 856

3 Disadvantages of sophistication 858

4 Resolute choice 860

5 Wise choice 863

6 But is wise choice really different from sophisticated choice? 868

Acknowledgments 872

References 873

36. The norms of practical reasoning 875

I Norms as categorical rational requirements 878

II Skepticism about the normativity of practical reasoning 882

III Contingency and context 884

References 891

Appendix: A Guide to Further Reading 893

Preliminary remarks 895

1 Further historical reading **897**
1.1 More readings on traditions and figures covered 897
1.2 Readings on traditions and figures not covered 898

2 Further reading at the intersection with action theory and moral psychology **902**
2.1 More readings on aspects of topics covered 902
2.2 Readings on topics not covered 903

3 Further reading on practical normativity **905**
3.1 More readings on aspects of topics covered 905
3.2 Readings on topics not covered 905

4 Further reading on practical rationality **909**
4.1 More readings on aspects of topics covered 909
4.2 Readings on topics not covered 909

References **911**

Index **924**

The Routledge Handbook of Practical Reason

Over the last several decades, questions about practical reason have come to occupy the center stage in ethics and metaethics. *The Routledge Handbook of Practical Reason* is an outstanding reference source to this exciting and distinctive subject area and is the first volume of its kind. Comprising thirty-six chapters by an international team of contributors, the Handbook provides a comprehensive overview of the field and is divided into five parts:

- Foundational Matters
- Practical Reason in the History of Philosophy
- The Philosophy of Practical Reason as Action Theory and Moral Psychology
- The Philosophy of Practical Reason as the Theory of Practical Normativity
- The Philosophy of Practical Reason as the Theory of Practical Rationality

The Handbook also includes two chapters by the late Derek Parfit, 'Objectivism about Reasons' and 'Normative Non-Naturalism.'

The Routledge Handbook of Practical Reason is essential reading for philosophy students and researchers in metaethics, philosophy of action, action theory, ethics, and the history of philosophy.

Ruth Chang is Chair and Professor of Jurisprudence at the University of Oxford and Professorial Fellow at University College, Oxford, UK.

Kurt Sylvan is Associate Professor of Philosophy at the University of Southampton, UK.

Routledge Handbooks in Philosophy

Routledge Handbooks in Philosophy are state-of-the-art surveys of emerging, newly refreshed, and important fields in philosophy, providing accessible yet thorough assessments of key problems, themes, thinkers, and recent developments in research.

All chapters for each volume are specially commissioned, and written by leading scholars in the field. Carefully edited and organized, *Routledge Handbooks in Philosophy* provide indispensable reference tools for students and researchers seeking a comprehensive overview of new and exciting topics in philosophy. They are also valuable teaching resources as accompaniments to textbooks, anthologies, and research-orientated publications.

Also available:

The Routledge Handbook of Dehumanization

Edited by Maria Kronfeldner

The Routledge Handbook of Anarchy and Anarchist Thought
Edited by Gary Chartier and Chad Van Schoelandt
The Routledge Handbook of the Philosophy of Engineering
Edited by Diane P. Michelfelder and Neelke Doorn
The Routledge Handbook of Modality
Edited by Otávio Bueno and Scott A. Shalkowski
The Routledge Handbook of Practical Reason
Edited by Ruth Chang and Kurt Sylvan
For more information about this series, please visit: www.routledge.com/Routledge-Handbooks-in-Philosophy/book-series/RHP

The Routledge Handbook of Practical Reason
Edited by Ruth Chang and Kurt Sylvan

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

First published 2021
by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN
and by Routledge
52 Vanderbilt Avenue, New York, NY 10017
Routledge is an imprint of the Taylor & Francis Group, an informa business
© 2021 selection and editorial matter, Ruth Chang and Kurt Sylvan; individual chapters, the contributors
The right of Ruth Chang and Kurt Sylvan to be identified as the authors of the editorial material, and of the authors for their individual chapters, has been asserted in accordance with sections 77 and 78 of the Copyright, Designs and Patents Act 1988.
All rights reserved. No part of this book may be reprinted or reproduced or utilised in any form or by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying and recording, or in any information storage or retrieval system, without permission in writing from the publishers.
Trademark notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.
British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library
Library of Congress Cataloging-in-Publication Data
Names: Sylvan, Kurt, editor. | Chang, Ruth, editor.
Title: The Routledge handbook of practical reason / edited by Kurt Sylvan and Ruth Chang.
Description: Abingdon, Oxon ; New York, NY : Routledge, 2021. | Series: Routledge handbooks in philosophy | Includes bibliographical references and index.
Identifiers: LCCN 2020038584 (print) | LCCN 2020038585 (ebook) | ISBN 9781138195929 (hbk) | ISBN 9780429266768 (ebk)
Subjects: LCSH: Practical reason.
Classification: LCC BC177 .R684 2021 (print) | LCC BC177 (ebook) | DDC 128/.33—dc23
LC record available at https://lccn.loc.gov/2020038584
LC ebook record available at https://lccn.loc.gov/2020038585
ISBN: 978-1-138-19592-9 (hbk)
ISBN: 978-0-429-26676-8 (ebk)
Typeset in Bembo
by Apex CoVantage, LLC

To the memory of Derek Parfit
and
To Teddy and Bertie, may you be ever more rational

Acknowledgments

This volume is the brainchild of one of us, who then approached the other of us, to help create the present volume. We both believed that it was time to collect together, name, and give an overview of the disparate and wide-ranging research of a program of philosophical inquiry that traverses ethics, metaethics, moral psychology, the philosophy of mind, action theory, epistemology, and beyond: hence *The Philosophy of Practical Reason*. Our aim is this volume is to lay out some of the main features of this terrain in a way that is both accessible to beginning students in the philosophy of practical reason and yet of interest to scholars working in the field.

We owe our greatest thanks to our contributors for sharing our vision, for making the volume what it is, and for their patience while we corralled, herded, and shepherded a large number of very busy people in order to bring the volume to life.

Without Rebecca Shillabeer at Routledge, this volume would not exist. We would also like to give special thanks to Adam Johnson at Routledge, who has been very supportive and helpful throughout the whole process. Thanks are also owed to three anonymous reviewers of the proposal for the volume, who suggested some important changes that helped to shape it. The year we submitted the volume saw the 50^{th} such volume in the *Routledge Handbook Series in Philosophy*, founded and overseen by Tony Bruce.

This volume is dedicated to Derek Parfit, who was one of the leading lights in the philosophy of practical reason (and so much more), an important mentor to both of us, and a dear friend of one of us. It is also dedicated to the small children of one of us, in the hopes that in their journey to adulthood, they will have many bouts of practical rationality.

Contributors

Elizabeth Anderson is John Dewey Distinguished University Professor of Philosophy and Women's Studies and Arthur F. Thurnau Professor at University of Michigan, Ann Arbor, USA.

Chrisoula Andreou is Professor of Philosophy at the University of Utah, USA.

Nomy Arpaly is Professor of Philosophy at Brown University, USA.

Michael E. Bratman is U. G. and Abbie Birch Durfee Professor in the School of Humanities and Sciences and Professor of Philosophy at Stanford University, USA.

John Broome is White's Professor of Moral Philosophy, *Emeritus*, at the University of Oxford, UK, and Honorary Professor at the Australian National University.

Sarah Buss is Professor of Philosophy at the University of Michigan, Ann Arbor, USA.

Agnes Callard is Associate Professor of Philosophy at the University of Chicago, USA.

Ruth Chang is Chair and Professor of Jurisprudence at the University of Oxford and Professorial Fellow at University College, Oxford, UK.

David Copp is Distinguished Professor of Philosophy, *Emeritus*, at the University of California, Davis, USA.

Jonathan Dancy is Professor of Philosophy at the University of Texas at Austin, USA.

Stephen Darwall is Andrew Downey Orrick Professor of Philosophy at Yale University, USA.

David Enoch is Rodney Blackman Chair in the Philosophy of Law at the Hebrew University of Jerusalem, Israel.

Patricia Greenspan is Professor of Philosophy at the University of Maryland, USA.

Elizabeth Harman is Laurance S. Rockefeller Professor of Philosophy and Human Values at Princeton University, USA.

Sally Haslanger is Ford Professor of Philosophy and Women's and Gender Studies at the Massachusetts Institute of Technology, USA.

Pamela Hieronymi is Professor of Philosophy at the University of California, Los Angeles, USA.

Margaret Olivia Little is Professor of Philosophy, Director of the Kennedy Institute of Ethics, and Co-Founder of the Ethics Lab at Georgetown University, USA.

Errol Lord is Associate Professor of Philosophy at the University of Pennsylvania, USA.

Coleen Macnamara is Associate Professor of Philosophy at the University of California, Riverside, USA.

Julia Markovits is Associate Professor of Philosophy at Cornell University, USA.

Jennifer M. Morton is Associate Professor of Philosophy at the University of North Carolina, Chapel Hill, USA.

Derek Parfit was Senior Research Fellow at All Souls College, Oxford, UK, and a visiting professor at Harvard University, New York University, and Rutgers University, USA. He died unexpectedly in 2017 before completing new work for this volume.

Sarah K. Paul is Associate Professor of Philosophy at NYU Abu Dhabi, United Arab Emirates.

Wlodek Rabinowicz is Professor of Philosophy, *Emeritus*, at Lund University, Sweden.

Peter Railton is Gregory S. Kavka Distinguished University Professor, John Stephenson Perrin Professor, and Arthur F. Thurnau Professor at the University of Michigan, Ann Arbor, USA.

Connie S. Rosati is Professor of Philosophy at the University of Texas, Austin, USA.

Geoffrey Sayre-McCord is Morehead-Cain Alumni Distinguished Professor of Philosophy at the University of North Carolina at Chapel Hill, USA.

T. M. Scanlon is Alford Professor of Natural Religion, Moral Philosophy, and Civil Polity, *Emeritus*, at Harvard University, USA.

Tamar Schapiro is Associate Professor of Philosophy at the Massachusetts Institute of Technology, USA.

Keshav Singh is Assistant Professor of Philosophy at Syracuse University, USA.

David Sobel is Irwin and Marjorie Guttag Professor of Ethics and Political Philosophy at Syracuse University, USA.

Kurt Sylvan is Associate Professor of Philosophy at the University of Southampton, UK.

Sergio Tenenbaum is Professor of Philosophy at the University of Toronto, Canada.

Valerie Tiberius is Paul W. Frenzel Chair in Liberal Arts and Professor of Philosophy at the University of Minnesota, USA.

Kenneth Walden is Associate Professor of Philosophy at Dartmouth College, USA.

R. Jay Wallace is Judy Chandler Webb Distinguished Chair for Innovative Teaching and Research and Professor of Philosophy at the University of California, Berkeley, USA.

Natalia Washington is Assistant Professor of Philosophy at the University of Utah, USA.

Monique Wonderly is Assistant Professor of Philosophy at the University of California, San Diego, USA.

David B. Wong is Susan Fox Beisher and George D. Beischer Distinguished Professor of Philosophy at Duke University, USA.

An introduction to the philosophy of practical reason

Kurt Sylvan and Ruth Chang

Over the last several decades, questions about practical reason have come to occupy center stage in ethics and metaethics. While such questions received considerable attention from some figures in the history of philosophy (most notably Aristotle, Hume, and Kant, whose ideas continue to shape contemporary work), philosophical reflection on practical reason took on a life of its own in the second half of the 20$^{\text{th}}$ century. This development is owed in large part to now-classic work on practical reason by Elizabeth Anscombe in the 1950s; Donald Davidson from the 1960s; Philippa Foot, Thomas Nagel, and Bernard Williams in the 1970s; Christine Korsgaard and John McDowell in the 1980s, 1990s, and 2000s; and Derek Parfit, T. M. Scanlon, John Broome, Jonathan Dancy, Michael Bratman, Michael Smith, and Joseph Raz in the 1980s through to the present.[1] The work of these figures stimulated research on many new issues concerning reasons, reasoning, the faculty of Reason, and rationality, issues which the current generation of thinkers now explores systematically in their own right. Arguably some of the most exciting work being done in ethics and metaethics today concerns these issues. This rich, diverse, and penetrating work gives rise to a distinctive area of inquiry that we propose to call the *philosophy of practical reason*.

The aim of this handbook is to provide a survey of research in the philosophy of practical reason, with some attention to the history of philosophy, but with an overall focus on the contemporary analytic tradition. We conceived of it as a *teaching* volume, something that might be suitable for advanced undergraduates and graduate students in philosophy, and each contribution has been written with that audience in mind. The volume fills a surprising lacuna in the literature: there has so far been no dedicated handbook on the philosophical study of practical reason.[2] Besides providing overviews of many central topics in the field, however, it also collects cutting-edge research drawn from both senior figures and younger scholars. Our approach has been to give the contributors a significant degree of freedom in pursuing their topics; we allowed them to work out new ideas rather than just assigning them topics and asking for guides to the existing literature. As a result, much new territory has been staked out in the course of providing a comprehensive map of a large and varied field.

Because of the vast scale of the territory surveyed within these pages, in this Introduction, we can give only a cursory overview of the field and describe the contributions of the authors within the context of that overview. An Appendix at the end of the book provides a guide to further reading with advice about where to look for specific topics not covered by this volume.

[1] See especially Anscombe (1957), Davidson (1963, 1970), Foot (1972a, 1972b), Nagel (1970, 1986), Williams (1979/1981), Korsgaard (1986, 1996a, 1996b, 1997, 2008, 2009a), McDowell (1978, 1979, 1995, 1998), Parfit (1997, 2001, 2011, 2017), Scanlon (1998, 2007), Broome (1999, 2004, 2005, 2007a, 2007b), Dancy (2000, 2004, 2018), Bratman (1987, 1999, 2018), Smith (1994, 1995, 2004), and Raz (2002, 2011).

[2] The closest volumes are Star (2018) and Mele and Rawling (2004). The first covers reasons and normativity in general (including epistemic and aesthetic normativity). The second covers rationality

in general; its date of publication means that it does not cover significant recent work. We recommend both as companions to our volume.

I What is the philosophy of practical reason?

The philosophy of practical reason has its roots in decades of scattered work originally written within more established areas of philosophy, especially action theory, ethics, metaethics, philosophy of mind, moral psychology, and the theory of rationality. The questions it asks – What is a reason? What is it to act for a reason? What is it be rational? to give just a few examples – beg for investigation across a range of traditional areas of philosophy. We suggest that the philosophy of practical reason is hence best characterized as a *question-driven* domain of inquiry.

In particular, we suggest that the field can be roughly characterized by its concern with questions belonging to one of three related branches of inquiry: i) the philosophy of practical reason as action theory, philosophy of mind, or moral psychology; ii) the philosophy of practical reason as metanormative or normative/ethical inquiry; and iii) the philosophy of practical reason as a theory of rationality. Carving up the domain in this way is in some ways arbitrary, but it lends a useful structure by which we can provide an overview of the subject, highlight many central issues within it, and systematize our authors' rich and diverse contributions. Our expectation is that as this fast-evolving field continues to develop, so too will salutary ways of conceptualizing and organizing new work within it.

1 The philosophy of practical reason as the philosophy of action, mind, or moral psychology

Suppose your nemesis has recently won an accolade and you send him your congratulations. Your friend asks you why you are congratulating someone who has been the bane of your existence. You say: 'It was the decent thing to do.' Here you are seeking to describe the reason *for which* you acted. The reason for which you acted is, to a first approximation, the consideration you would cite in a cool moment when asked why you did something. It explains, by your lights, why you did what you did, along with what you would consider a justification for your action. Sometimes, however, the consideration that you believe motivated you to act may not be what in fact motivated your action. You might, for example, later go to your therapist and describe your mixed feelings about the situation, and your therapist might then tell you that

although you believed you were motivated by decency, in fact you were motivated by your desire to be liked, even by someone who has treated you badly. Both types of reasons are often lumped together as 'motivating reasons,' reasons that figure in rationalizing explanations of why you did what you did.

When practical reasons are understood as motivating reasons, the study of practical reason becomes the study of the objects and operations of a *mental faculty*; these include reasoning, intending, acting, and desiring for reasons. Hence this first branch of the philosophy of practical reason interacts heavily with the philosophy of mind, moral psychology, and action theory, pondering questions such as:

- What is the nature of motivating reasons? Are they psychological states? If so, which ones are fundamental? If not, what else could motivating reasons be?
- What is practical reasoning, and how is it related to theoretical reasoning? Is it even possible for 'reason' to be practical?
- What happens when someone acts?
- What is the relationship between practical reasons, intentional action, and autonomy?

While the pre-20$^{\text{th}}$ century history of philosophy contained important answers to some of these questions, focused research into them took off in the mid-20$^{\text{th}}$ century, owing to the influence of Elizabeth Anscombe (especially her 1957 book *Intention*) and Donald Davidson (especially his 1963 paper 'Actions, Reasons, and Causes'). These two figures had importantly different visions of the proper explanation of rational action. Anscombe argued that there is a distinctive sense in which we can ask 'why' someone did something and that this 'why' question could be properly answered only by citing reasons, not causes. Davidson, in response, defended with vigor a causal approach to reasons-explanations of action. Two traditions emerged from their classic work, which continue to enjoy adherents today, with figures like Sebastian Rödl (2007), Michael Thompson (2008), and Candace Vogler (2002) developing views that vindicate different themes from Anscombe and a long list of researchers in philosophy of mind either defending Davidson's 'belief-desire' action theory (see, e.g., Dretske 1988, Mele 1992, 2003 and Sinhababu 2017) or extending and reworking it in key ways (see, e.g., Bratman 1987).

Other controversies in this area have older roots. Another debate, related to the previous one but worth distinguishing, is the debate between Humeans and anti-Humeans about motivation.[1] For Humeans, *desire* is the fundamental motivating state. Humeans accept the Davidsonian claim that desires need to be informed by beliefs in order to

[1] This debate should be distinguished from a different Humeanism/anti-Humeanism debate about *normative* reasons that we will discuss in the next section.

produce actions, but they think that only desires can be *intrinsically* motivating. Anti-Humeans reject these claims. At a minimum, they claim that beliefs can be motivating states in their own right and that motivation may not require desire in a sense that can be identified independently of the agent's beliefs about what ought to be done (Nagel 1970). Anti-Humeans need not side with Anscombe about reasons-explanations: they can allow that action remains grounded in a causal relation between the agent's beliefs and her bodily movements. But Anti-Humeans can agree with Anscombe that reasons-explanations are different in kind from causal explanations. Kant, for instance, treated desires as parts of the natural order, but he thought that practical reason stood at a reflective distance from the natural order, operating autonomously to produce full-blooded action. This picture of motivation – as well as a further Kantian view about normativity – has become associated with Korsgaard, who influentially defended it in her 1986 and 1996b works.

In addition to these two long-standing debates, new paradigms have emerged in the first branch of the philosophy of practical reason. For example, after Dancy's (2000) opposition to the psychologism about motivating reasons assumed by the Davidsonian tradition, many philosophers of practical reason have been converted to the view that motivating reasons are non-psychological entities such as facts, propositions, or states of affairs. To take another example, there have recently been revivals of the topic of practical reasoning by Richardson (1994), Millgram (2001), Wallace (2001), and Broome (2013). Some related areas of recent activity concern 1) the relationship between practical and theoretical reasoning (see Wallace 2001, Setiya 2007, Bratman 2009a, and Dancy 2018) and the question of whether acting for a reason is, in part or in whole, an *intellectual* achievement (see Fix 2018 for a pathbreaking discussion) and 2) the nature and features of the faculty of practical reason (see Raz 2002, 2011 and Korsgaard 2008, 2009a, 2009b). There is much else in this first branch that we do not try to cover in this volume, though the Appendix provides a guide to further literature, and we will return to these issues in describing the contributions to the volume in Part II.

2 The philosophy of practical reason as metaethics and ethics

In contrast to the concept of a motivating practical reason, there is the concept of a *normative* practical reason. This concept enables us to investigate the reasons that *count in favor* of acts and motivational attitudes like desire and intention. Suppose your dentist tells you that the soft tissue inside your root canal is inflamed and that you will need to have a root canal procedure. The inflammation of your soft tissue is a normative reason for you to have the operation – it counts in favor of your having it.

Normative reasons justify our actions. They are also in play when we are open to criticism for failing to act in certain ways (e.g., for not getting a root canal operation). Although normative reasons should be distinguished from motivating reasons, it is sometimes true that we are motivated by normative reasons. If you are motivated by the normative reasons you possess and do what they favor doing, you are *substantively rational* – you have recognized and responded to your normative reasons.

Normative reasons are determinants of how we should live. The second branch of the philosophy of practical reason is dedicated to the study of practical reasons in this sense. This branch interacts heavily with ethics and metaethics, examining questions such as the following:

- Metaphysical questions:
- What is the nature of normative practical reasons? What makes something a normative reason? What, exactly, is normativity? What is the role of agency in understanding normative reasons? Can normative reasons be privately held? — What is the relationship between normative practical reasons and other normative practical phenomena such as value and obligation? Do reasons explain values and obligations? Or are reasons explained in terms of more fundamental normative phenomena? — How do normative reasons justify action? What are the features of choice contexts through which normative reasons determine what we should do?
- High-level substantive questions:
- Are normative practical reasons explained by principles? Or are reasons holistic in a way that undermines principles? — Do reasons have different kinds of normative weight? For example, might normative reasons have both requiring weight and recommending weight? How can different kinds of reasons be 'put together' to yield all-things-considered conclusions about what one should do? — Is choice determined by the balance of objective, value-based reasons, or is there underdetermination in what these reasons demand, creating space for the will or the self to make reasons of its own?

Although the first branch we discussed was also the first to be explored systematically in the 20$^{\text{th}}$ century, it is really the growth of this second branch beginning in the 1970s and exploding in the 1990s and 2000s that established the philosophy of practical reason as a central area of philosophical inquiry in its own right and indeed among the most prominent of 21$^{\text{st}}$ century research areas.

There is an interesting story behind this rise, which involves ethics being virtually colonized by work belonging to this second branch, so that today, research on practical reason assumes centrality in both of the traditional fields of ethics and metaethics. Beginning in the 1970s, ethics and metaethics came to be framed in the ideology of

reasons by many philosophers. Initially, the relationship between ethics and practical reason was fraught: an appeal to a broadly Humean view of normative reasons was used to question the normativity of morality by Foot (1972a), Mackie (1977), and Williams (1979). While Korsgaard (1986, 1996a, 1996b) and Smith (1994, 1995) pushed back against Humean accounts of normativity and defended the categorical normativity of morality, both took practical reason to have explanatory priority over morality.[2] But a key reversal took place beginning in the late 1990s with the work of Parfit (1997) and Scanlon (1998). Their commonsense methodology led back to the view that ethics is an evident source of practical reasons in its own right, a view which had not been so dominant since the intuitionist era of British moral philosophy from Sidgwick to Ewing. This methodology has been accepted even by philosophers who believe – *contra* Parfit and Scanlon – that normative reasons have as their source not only ethical facts but also volitional activities such as commitments. *Hybrid voluntarists* such as Chang (2009, 2013a, 2013b, 2017), for example, accept both the objective moral reasons of common sense and reasons grounded in commitments of the self.

As this little story suggests, a central controversy in the second branch concerns the relationship between normative reasons and motivation. One can usefully frame this debate as a debate about whether normative reasons can be explained (partially or fully) in terms of actual or counterfactual facts about agents' desires or other motivating states.[3] Accordingly, some Humeans (e.g., Schroeder 2007) claim that the fact that there is a normative reason for an agent A to φ is explained by the fact that φ-ing would help to fulfill some of A's desires; other Humeans (e.g., Brandt 1979) claim that the fact that there is a normative reason for A to φ is explained by the fact that φ-ing would promote certain *idealized* desires that A *would* have under certain conditions. There is, then, a variety of ways of rejecting Humeanism. In starkest opposition are Parfit and Scanlon, who deny that *any* facts about normative reasons are explained by motivating states and indeed think that there can be no metaphysical explanation of normative reasons in non-normative terms. But there are other non-Humeans who

[2] There were two other prominent contributors between the 1970s and 1990s – Jonathan Dancy (1993) and John McDowell (1978, 1995, 1998) – who rejected the Humean approach without accepting anything like the Kantian internalism of Korsgaard and Smith, with Dancy favoring an intuitionist, non-naturalist realism (see Dancy 2006) and McDowell favoring a virtue-based, Aristotelian approach (see 'Virtue and Reason' in McDowell 1998). Note also that Foot eventually abandoned the apparently Humean view in her earlier work with a turn to an Aristotelian approach in Foot (1978, 2001).

[3] There are several different kinds of 'explanation' that might be given here, it is worth noting. Much recent literature focuses on metaphysical explanations: this literature takes it for granted that there are *facts* about reasons, and the question is to understand whether and how these facts might be *grounded* or *analyzed* in terms of facts about motivation. But there are other explanations one could seek and that some theorists do seek. One might, for example, seek a *normative* explanation of facts about normative reasons in terms of facts about desires. Alternatively, one might engage in conceptual or linguistic ascent and examine whether the *concept* of a normative reason or the *word* 'reason' used normatively can be *conceptually* or *semantically* explained in terms of concepts or language picking out or expressing motivating states.

are less starkly opposed or who at least have a related explanatory agenda. Hybrid voluntarists, for example, will allow that *some* normative reasons are explained by internal volitional states. Some Kantians (e.g., Korsgaard) will agree with Parfit and Scanlon that mere desires do not generate normative reasons but will not follow that pair in holding that normative reasons lack further metanormative explanation: instead, they will seek to explain normativity by appealing to the constitutive principles of the mental faculty of practical reason. Hence Kantians and Humeans exemplify a broader approach sometimes called *internalism* about normative reasons, which seeks to understand normative reasons in terms of mental phenomena or their constitutive normativity.[4]

It is worth remarking briefly on the arguments that structure the Humeanism/anti-Humeanism controversy in the second branch. Perhaps the central argument for Humeanism about normative reasons is an argument from *naturalism*: according to Schroeder (2007), for example, this view provides the best account of how the normative is grounded in the natural and hence is essential for securing the view that nature is all that fundamentally exists. Opponents of Humeanism, by contrast, often suggest that it pays far too high a price for its naturalistic credentials. In particular, both non-naturalists like Parfit and Scanlon and Kantians like Korsgaard argue that Humeanism fails to explain the reason-giving power of morality and even the reason-giving power of *prudence* (or *objective self-interest*). While some Humeans claim that they can vindicate our intuitions about the normativity of morality and prudence (see Schroeder 2007), others are willing to pay the price: they *explain away* some intuitions (see, e.g., Street 2009) or draw attention to opposing aspects of commonsense thought (see, e.g., Manne 2014), thereby returning to a view held earlier by Foot (1972a, 1972b). Hearkening back to Mackie (1977), Humeans may also push back against the alleged intuitive credentials of non-naturalism, noting that any commonsense picture of reality will include no mysterious third realm populated by insubstantial normative truths. Kantians then seek to stand above this fray, arguing that we can vindicate both case-based intuitions and commonsense metaphysics without being Humeans: the

[4] These approaches are also sometimes put under the umbrella of *constructivism* about normative reasons, since they can be viewed as seeking to build facts about reasons out of facts about either desires or the constitutive principles of the faculty of practical reason; see Street (2008) for this way of presenting the terrain. Not all Kantians, however, see their project as constructivist (or at least as *automatically* constructivist); see, for example, Skorupski (2010, 2017) and Markovits (2014) and Markovits's discussion of constructivism with Walden in this volume.

There are other schemes of classification worth knowing about. Sometimes the debate between Humeans and Parfit-style anti-Humeans is examined under the heading of *Subjectivism* vs. *Objectivism*. But this alternative scheme of classification is not always used to distinguish between different views about the *metaphysical priority* between reasons and motivation. It is sometimes used to identify different views about what kinds of things can *give* or *be* reasons. Hence some Subjectivists – for example, Sobel in his chapter – are primarily interested in arguing that reasons are all ultimately *given* by subjective states rather than objective features of the world. They may then leave open whether the reason-for relation can be metaphysically analyzed.

constitutive principles of practical reason have categorical authority deriving from our distinctive nature as autonomous beings.

While the debate about the relationship between reasons and motivation is especially prominent in the literature, there are other angles from which the second branch studies normative reasons. Many theorists have been interested in the relationship between normative reasons and other normative categories, such as *value* and *obligation*. Here one important debate is between the *Reasons First* approach to normativity,[5] which seeks to understand all normativity in terms of reasons, and alternatives that either prioritize some other normative category (e.g., value) or deny priority to any category. This debate crosscuts the controversies about normativity and motivation. Hence Reasons Firsters include both Humeans like Schroeder as well as anti-Humeans like Scanlon. Besides this debate, recent theorists have also been interested in other metaphysical questions about normative reasons, investigating, for example, what kinds of things can stand in reason-relations (e.g., states of affairs, propositions, or mental states).

While the most striking divisions in the literature are along broadly metaphysical lines, there are other important divisions in the first branch that instead reflect different answers to more abstract first-order questions about practical reasons. Some of these divisions are connected to familiar ones in moral philosophy (most centrally, consequentialism vs. deontology vs. virtue ethics).[6] But many crosscut familiar divisions in first-order ethical theory. There are enough issues under this heading to justify their own handbook. But a few important divisions reflect different answers to questions about:

1. the relationship between reasons and *principles*;
2. how to understand the *weight* and *force* of normative practical reasons;
3. the status of *pluralism* about normative practical reasons and practical 'oughts';
4. the *determinacy* of practical reasons and their *comparability*;

Under heading 1 is the dispute between *particularists* about reasons, like Dancy (2004), who deny that reasons are underwritten by principles, and *principled* theorists like Kant and Scanlon. A related issue under heading 2 is whether reasons have 'atomistic' weights – that is, each reason can be assigned a fixed weight so that what there is

[5] 'Reasons First' is sometimes used to refer to a stronger view held by Parfit and Scanlon, according to which reasons i) explain all normativity and ii) admit of no further explanation in naturalistic terms. But owing to the rise of views like Mark Schroeder's, 'Reasons First' has recently been used to refer only to the view that reasons are basic within the normative domain (where it is left open whether they admit of naturalistic grounding).

[6] See Korsgaard (1996b) and Schapiro (2001) for rich, historically grounded discussions of the relationship between conceptions of agency and moral theories.

overall reason to do is determined by balancing these weights against each other – or whether the weight of a reason is a *holistic* matter (*atomism vs. holism about reasons*).[7]

A different issue under heading 2 concerns the kind(s) of *force* that normative reasons have. Scanlon (1998) had influentially suggested that reasons are considerations that *count in favor* of acts and attitudes. The notion of favoring seems to contrast with what Dancy (2004) called *peremptory* normative concepts like *obligation* and *requirement*. It is unclear how to explain one kind of normative force in terms of the other, raising the question of whether it might be better to think of reasons as having two kinds of normative force, as Gert (2007) has suggested. Whether we should follow Gert in treating this distinction as basic, and how more generally we should understand the distinction, are important questions in the background of several debates in the philosophy of practical reason.

In addition to examining whether there are fundamentally different kinds of normative *force*, we might also consider whether there are fundamentally different *flavors* of practical normativity. We might, for example, wonder whether the distinction between prudential and moral reasons is a fundamental joint in practical reality; we might also wonder about whether there other kinds of reasons for action connected with non-moral and non-prudential values (e.g., reasons for action grounded in aesthetic value). If we follow a commonsense approach and treat all the intuitively significant factors that weigh with us as good candidates for being genuine reasons, it may appear that pluralism is forced upon us. Yet it may also seem there must be an overarching standard of comparison if we are to reach verdicts about what we have *most reason* to do. These puzzles about the plurality and comparability of practical reasons generate further important fault lines in the literature. For a stark contrast on these issues, one can compare Copp's contribution to this volume with Chang (2004a, 2004b, 2015).

Even if reasons can always be compared, we are left with the question of whether reason can always reach action-guiding verdicts. Some have taken pluralism to expose the limits of practical reason and the indeterminacy of practical obligation. Perhaps standard forms of rational choice theory cannot make room for 'transformative' choices (Paul 2014), or any rational theory must make room for our responding to vaguely formed, inchoate, 'proleptic' reasons (Callard 2018). Others have rejected pluralism precisely because of their optimism about reason's power to determine action-guiding verdicts. Yet others believe that pluralism is no threat to the determinacy of practical reason. And still others – including one of us – have suggested that determinate verdicts can sometimes only be achieved by the intervention of the will: to resolve a hard choice, one must *create* new practical reasons. These disagreements have played out for several decades and continue to represent hotspots of research.[8]

[7] For some important discussions after Dancy (2004), see Schroeder (2011) and the Introduction and papers in Lord and Maguire (2016).

[8] For an early effort to take stock and collect work on these issues as well as the issues described in the previous paragraph, see Chang (1997).

3 The philosophy of practical reason as theory of rationality

Practical reason can be approached from a third angle. Suppose you are engaged in a game of chess. Your opponent has just put you in check. It is natural to say here that you have a reason to move out of check. You are justified in moving out of check but not in the same sense in which you are justified in having a root canal. Your 'justification' for moving your king is *relativized* to the practice of chess. Now consider instead the activity of thinking or deliberating about what to do. Like chess, this activity has certain associated rules and standards. When your practical deliberation is governed according to these rules of the game, we can call it 'structurally rational' (Scanlon 2007). If you want to kill your enemy and believe that poison will do the trick, then you are required by structural rationality to form an intention to get some poison (or give up one of your other attitudes). While you may be 'rationally required' to get the poison, one might think that you don't have a normative reason to get some, since poisoning your enemy is not something that there is any good reason to do.

The third branch of the philosophy of practical reason is dedicated to questions about *practical rationality* in the structural, not substantive, sense. Recall that you are substantively rational if you recognize and respond to your normative reasons. By contrast, you are structurally rational if you follow the rules governing movements of your mind, from one attitude to another, and from certain attitudes, like intention, to action. This branch interacts heavily with philosophical work on the mental faculty of rationality (including its epistemic side), and some of the questions it investigates include:

- What are the requirements of structural rationality? Is there any normative reason to obey these requirements?
- Are the requirements of structural *practical* rationality derived from those of structural *epistemic* rationality? Which requirements are the most basic? Can all requirements be explained in terms of instrumental rationality?
- Is being moral a rational requirement?

Scanlon (1998: Ch.1) argued that challenges to the idea that morality as a source of normative reasons tend to rest on a confusion between what is a good reason and what is required by rationality.[9] Morality may be a source of good reasons even if we are not rationally required to be moral or irrational for being immoral. While this distinction between good reasons and rationality may initially sound surprising, it can be easily appreciated by reflecting on examples from other normative domains. Consider the

[9] He was echoing a thought that had appeared earlier in McDowell's (1978, 1995) responses to Foot and Williams.

epistemic domain. If there exists a proof of some theorem from some true axioms, there is a clear sense in which there is a *conclusive reason* to believe the theorem. But if the proof is sufficiently complex and the theorem is sufficiently unobvious, there may be nothing *irrational* in failing to believe the theorem even if one believes the axioms.

Around the same time Scanlon made this point, a related distinction was drawn by John Broome (1999), who separated rationality as a source of *coherence requirements* on sets of attitudes and responsiveness to reasons for particular attitudes. Coherence requirements include what Broome calls "Enkrasia," which requires one to avoid the akratic combination of believing that one ought to while failing to intend to . He also applied this approach to explain the *instrumental* irrationality involved in failing to intend to take what one believes to be the necessary means to one's ends, and was followed by Wallace (2001), Way (2010), and others.[10] While Kolodny (2005) challenged Broome's claim that rationality only requires certain *combinations* of attitudes and argued that rationality requires one to resolve incoherence in specific ways, he maintained the Scanlonian distinction between questions of reasons and questions of rationality; indeed, he called the normative authority of rationality into question (see also Kolodny 2007). The importance of the Broome–Scanlon distinction was also recognized by several other influential philosophers of practical reason (see Dancy 2000, Parfit 2001, 2011, Raz 2005, 2011, and Bratman 2009b, 2018), which helped to hasten the separation of research on rational requirements and research on normative practical reasons. While some have recently argued that the intuitive distinction between reasons and rationality doesn't motivate free-floating coherence requirements (see Kiesewetter 2017 and Lord 2018), the independent study of rationality remains alive and well.

There are many other issues that we would include within the philosophy of practical reason that are exciting. For further information about these and other issues, the reader may consult the Appendix to this volume.

[10] While Broome is usually credited with the idea that rational requirements are at bottom coherence requirements on *sets* of attitudes rather than requirements to adopt particular attitudes, it can actually be traced back to Hill (1973), Greenspan (1975), and Dancy (1977) (who merely reaffirmed it in Dancy 2000).

II A guide to the volume

We have structured the volume in light of the foregoing divisions of the field. To help the reader, we will now walk through the volume and explain how to place the contributions in the map of the field just drawn.

1 Foundational matters

The first part of the volume offers some big-picture reflections on what the philosophy of practical reason should be about and on how best to answer its central questions. The main aim of T. M. Scanlon's chapter is to call attention to seven questions that he thinks any student of the philosophy of practical reason should think about and try to answer and to call attention to some key presuppositions of objectivist and subjectivist answers to these questions. Along the way, he also states and gives a brief defense of some of his own views. He begins with a brief defense of the cognitivist and realist approach he has long defended (see Scanlon 2013). He then defends a distinction between reasons and rationality and uses this distinction to cast doubt on approaches that put rationality before reasons. Broome's chapter defends his version of the distinction between what reasons require and what rationality requires. He argues that normativity and rationality can only be brought together via a Kantian conception of rationality that he suggests is "far from our ordinary concept of rationality."

Peter Railton asks how it is possible for reason to be practical. He notes that there are two kinds of answers, guided by narrower and broader understandings of the phrase 'practical reason.' The narrow conception would seek to answer the question by showing how there is a distinctive form of *reasoning* which is practical; the broader conception would seek to answer it by showing how there is a *set of capacities* which work together to enable one to respond to normative reasons for action. Railton draws on the neglected areas of overlap between the Aristotelian, Kantian, and Humean traditions to defend a novel view about how reason (understood broadly) could be practical; he ends by suggesting that this view is confirmed by the empirical study of motivation and action.

The next three chapters, by Sally Haslanger, Elizabeth Anderson, and Ruth Chang, offer alternative conceptions of the field. Scanlon, Broome, Railton, and many others bracket the social role of practical reason and the context in which it occurs. Haslanger's chapter pushes back against this kind of bracketing, noting that it represents a kind of ideal theory. She argues that creatures who can respond to reasons as dominant

approaches already come to the table with certain social capacities and suggests that "[t]hese more basic forms of sociality are where we might find the sources of our practical orientations; they are the social preconditions for much of our thinking and acting." She then explores some ways in which practical reason can be understood to be socially and culturally conditioned, drawing on non-ideal theory in social and political philosophy. She ends with an alternative vision of the field. Anderson gives a more specific example of how to integrate philosophy of practical reason with non-ideal theory. She defends a *pragmatist* approach which conceives of normative judgments as tools for solving practical problems, tools which can be sharpened by experimentation and engagement with empirical data about the biases that shape our reasoning. Chang takes as her target the common idea, accepted by Scanlon, Broome, Railton, and many others, that being practically (substantively) rational is largely a matter of recognizing and responding to reasons. She suggests that this understanding of rational agency is too passive and suggests a more 'activist' view of what it is to be a rational agent. According to such activist views, agents have the normative power to *create* reasons and thus to determine which reasons they have. Chang urges that such an activist picture is needed to explain how we can pay more than just lip service to the idea that we are the authors of our own lives.

2 Practical reason in the history of philosophy

The second part of the volume examines some figures and traditions that have contributed importantly to the study of practical reason. As the chapter by David Wong illustrates, philosophical reflection on practical reason did not begin in the West. Wong surveys some relevant work in Chinese philosophy that was produced as early as the 6$^{\text{th}}$ century BCE from the Confucian and Daoist traditions. Both illustrate in different ways an overall approach that is *particularist*, *intuitionist*, and *virtue-theoretic*; this fact undermines a narrative that would trace such ideas to Aristotle (4$^{\text{th}}$ century BCE).

The volume then travels forward in time and westward in space to ancient Greece, with a chapter by Agnes Callard on some of Aristotle's contributions to the study of practical reason. Callard focuses on Aristotle's account of practical deliberation. She argues that Aristotle had a *geometrical* model of practical deliberation: the agent begins deliberation with a fixed end and then works backwards to derive an action appropriate to this end in a way inspired by geometrical analysis. Along the way, she contrasts this model with *evaluative* models of deliberation from elsewhere in historical and contemporary philosophy.

Tamar Schapiro's chapter considers an evaluative approach from Kant which contrasts with the non-evaluative one Callard finds in Aristotle (though Schapiro's main foil for the Kantian approach is a *mechanistic* approach she locates in figures as otherwise different as Leibniz, Davidson, and Bratman). She argues that the difference

between Kant's approach and mechanistic approaches owes to deeper difference in method, reflecting two conceptions of the purpose of the philosophy of practical reason. As Schapiro puts it, the point of a philosophy of practical reason according to the mechanistic tradition is to "explain what happens when someone acts," whereas, according to the Kantian tradition, "its aim is to show us what we are doing insofar as we are acting."

Geoffrey Sayre-McCord then gives a striking reading of Hume on practical reason. While Hume is universally acknowledged as one of the most important philosophers to write about practical reason, he is often interpreted as placing severe limits on reason's capacity to guide action, either as denying outright that reason can be practical or as advocating an instrumentalist picture on which reason's sole practical role is the coordination of means and ends. Against these interpretations, Sayre-McCord argues that Hume had a *robust* theory of practical reason: he allowed that reason can be practical in its own right and took its exercise to go beyond ensuring means-end coherence. Indeed, according to Sayre-McCord, Hume "makes important room for our deliberating about what to do specifically in terms of what is right, permissible, valuable, or virtuous and then acting accordingly as a result."

The section closes in the mid-20$^{\text{th}}$ century, with a chapter by Keshav Singh on insights into practical reason from a book that Davidson described as the most important work in the theory of action since Aristotle, Anscombe's *Intention*. Singh begins by observing that Anscombe's book has been much better appreciated in action theory than in the philosophy of practical reason. But he suggests that its lessons for the philosophy of practical reason are equally significant. Besides showing that Anscombe anticipated the *non-psychological* ontology of reasons associated with Jonathan Dancy, Singh argues that Anscombe's non-causalist account of acting for a reason remains a worthy solution to the problem of deviant causal chains and merits reconsideration in the general theory of reasons and rationality, especially given the non-causalist turn that has been independently taken in recent literature on the epistemic basing relation.

3 Practical reason, action theory, and moral psychology

The third section covers the first branch of the philosophy of practical reason. It focuses on:

1. the relationship between intentional action, acting for reasons, and deliberation, and intersecting issues about the relationship between autonomy and reasons-responsiveness (Arpaly and Buss);

2. the nature of practical reasoning and its differences and similarities to theoretical reasoning (Dancy);

3. the role of normative beliefs/appearances in intentional action and the relationship between beliefs about reasons and motivation (Tenenbaum and Rosati);

4. the role of emotions in practical reasoning (Greenspan);

5. the intersection of the first branch and relevant empirical work (Wonderly and Tiberius and Washington).

The section opens with a chapter by Nomy Arpaly which seeks to disabuse the reader of some doctrines about the relationship between agency and reflective or deliberative reasons-responsiveness.[1] In particular, she seeks to debunk views which treat a person's behavior as *more agential* in virtue of being *guided by deliberation* or in virtue of being *reflectively endorsed*. She argues that the basic case of acting for reasons is *unreflective*. Arpaly's contribution is followed by a chapter by Sarah Buss which considers the possibility of what she calls "passive agency." Against the view widely shared by philosophers and nonphilosophers alike, Buss argues that we cannot wittingly defy our own normative verdicts. Nonetheless, she argues, our reasoning selves can be dissociated from our acting selves, and this means that we can be passive bystanders to our own actions. Having reviewed the different forms that such dissociation can take, Buss concludes that, because "the capacity to reason is not the capacity to eliminate every element of arbitrariness from one's actions ... some measure of passivity is ... a necessary condition of everything that we do."

We then shift to a multifaceted chapter by Jonathan Dancy. Dancy suggests that two views that he has long defended – holism about normative reasons and non-psychologism about motivating reasons – can be used to clear space for a view about practical reasoning that he has recently adopted (see Dancy 2018). The new view that Dancy connects to his earlier views is the Aristotelian view that action is the proper conclusion of practical reasoning. Dancy argues that once we have the right views about normative and motivating reasons, nothing stands in the way of this Aristotelian view. Once this space is cleared, we can adopt a simple picture of the difference between practical and theoretical reasoning: practical reasoning is reasoning which properly concludes in action, while theoretical reasoning is reasoning which properly concludes in belief. In addition to providing a concise defense of this picture, Dancy's chapter also serves to acquaint the reader with his earlier influential work on normative and motivating reasons.

The next two chapters concern the relationship between normative beliefs/appearances, action, and motivation. Sergio Tenenbaum's chapter discusses the ancient doctrine that intentional action takes place 'under the guise of the good' – that is, the view that if an agent X is to do some act A intentionally, it must appear to X that there is something in favor of A-ing. The piece introduces the reader to the main arguments

[1] For earlier influential works in this vein, see especially Arpaly (2000, 2003) and Arpaly and Schroeder (2012).

for and against this view, which Tenenbaum has defended at great length elsewhere (see especially Tenenbaum 2007). Connie Rosati's chapter takes on a debate about a different alleged connection between normative beliefs and agency: the *judgment internalist* or *motivational internalist* view that if an agent X believes that there is good reason to do some act A, then X must be motivated to do A.[2] Her chapter surveys the arguments for and against this view (which she has previously evaluated alongside some related doctrines in Rosati 2016).

Then we have a chapter from Patricia Greenspan on the role of emotion in practical reasoning. Greenspan has long opposed the view that emotion's influence on action is entirely non-rational (1988). Her main ambition in this chapter is to review, update, and correct some misunderstandings of her work. Throughout, Greenspan argues that emotions play a normative role in practical reasoning, by supplementing and sometimes substituting for evaluative judgments.

The final two chapters address some interactions between empirical research and the first branch of the philosophy of practical reason. The first, by Monique Wonderly, examines the lessons that research on the nature of psychopathy provides for the philosophy of practical reason. A central lesson she draws from the study of psychopathy is that "practical reason is not a unary capacity but involves a suite of abilities that engage different aspects of our psychology and work together to help constitute us as unified agents." For, as Wonderly explains, psychopaths exhibit a surprising combination of excellence in some forms of practical reasoning and incapacity in others. Valerie Tiberius and Natalia Washington follow up Wonderly's case study by turning a wide-angle lens on the implications of social scientific research for the philosophy of practical reason. As they note, while there has been a wave of work by figures like Greene, Prinz, Nichols, Doris, and others on how moral psychology could be informed by social science, considerably less work has been done on practical reason. They pave the way for further research by considering possible lessons from social science for three topics: the moral rationalism vs. sentimentalism debate, the status of the link between intentional action and reasons, and an ameliorative approach to practical reasoning. They offer a balanced assessment of the bearing of social scientific research, noting that while approaches to practical reason often make empirical assumptions that demand scientific scrutiny, one must more careful about spotting these assumptions than some have been in the literature on moral psychology.

4 Practical reason and normativity

The fourth section turns to practical normativity. It is divided into two parts, one on the nature of normative practical reasons and another on high-level first-order

[2] This view must not be confused with a different view about *normative* reasons called 'internalism' (i.e., the view that if one has a normative reason to A, one must be motivated to A). Darwall (1983) called this view *existence internalism* and the view at issue in Rosati's piece *judgment internalism*; the

questions about such reasons. The first section contains chapters on objectivism vs. subjectivism about normative reasons (Parfit and Sobel), Kantian constructivism and constitutivism (Markovits and Walden and Enoch), and non-naturalism (Parfit). The second section contains work on the different types of force that normative reasons can exhibit (Little and Macnamara and Wallace), the status of pluralism about normative reasons and the question of whether there is a bare practical 'ought' or only moral and prudential 'ought's (Copp and Harman), and the nature of distinctively moral reasons for action and intention (Darwall).

This section could have naturally housed some other contributions in the volume. For example, we might have naturally enough placed Railton's, Scanlon's, and Chang's chapters in the first subsection; Railton's chapter is a new installment in the series of naturalistic realist works he has produced over the decades (see Railton 1986 for the *locus classicus*); Scanlon's chapter stands up for the non-naturalist cognitivism associated with him and Parfit; and Chang proposes a nonstandard, 'will-based' view of the grounds of practical reasons. We might also have placed Dancy's chapter here, since it is intended to be a new extension of his approach to the metaphysics of reasons. Hence the reader should consider reading these chapters along with this section.

a The nature of normative practical reasons

The first part opens with a chapter from Derek Parfit, to whom the volume is dedicated. He had generously agreed to write a new chapter for the volume on non-naturalist cognitivism, but he sadly died unexpectedly before he could generate something to print. We were therefore given unusual permissions from Routledge and Oxford University Press to keep a part of him in the volume through a reprint of some selections from *On What Matters* that we think all students of ethics and practical reason should read. These chapters bookend the subsection and feature accompanying text by the one of us who knew him best.

When investigating the nature of normative reasons, we should distinguish the question of *which* sorts of considerations can *be* normative reasons from the question of what *makes* such considerations normative reasons. To keep these questions apart, we recommend that 'objectivism' be used to denote the view that the kinds of considerations that can be reasons are *objects* of our desires and aims and 'subjectivism' be used to denote the view that the only kinds of considerations that can be reasons are facts about what would fulfill those desires. 'Objectivism' and 'subjectivism,' then, are contrasting views about what things could play the role of being normative reasons. Other 'isms,' such as 'nonnaturalism,' 'constitutivism,' and 'naturalism,' are views about the metanormative grounds of our reasons, not about which kinds of considerations can *be* reasons.

---

latter view is also commonly called *motivational internalism* (hence Rosati's title). Rosati has elsewhere discussed a relative of the former view about what is *good for a person* (1996).

These two questions are often combined as parts of broader accounts of the nature of normative reasons. The first two chapters of this section are no exception. Parfit's chapter defends *objectivism* about normative practical reasons, which he characterizes as the view that all normative practical reasons are given by features of the *objects* of our desires and aims. According to objectivism, these features also justify the desires and aims themselves and thus are reasons to have them. Objectivism is contrasted with subjectivism, which he characterizes as the view that "our reasons for acting are all provided by, or depend upon, certain facts about what would fulfill or achieve our present desires or aims." Parfit defends objectivism by opposing subjectivism and in particular by attacking it in the cases in which it might have appeared strongest. It would seem that subjectivists would have an easy time of explaining *prudential* as opposed to *moral* reasons, but Parfit offers a simple and now well-known argument to the contrary – the 'Agony Argument.'

One of the most important critics of that argument – and one of the most important defenders of subjectivism – is David Sobel.[3] Sobel outlines a strategy of argument for subjectivism that has three stages: offense, non-moral defense, and moral defense. The first stage draws attention to cases where subjectivism seems most intuitive, which tend to be cases of matters of mere taste in which what the agent, for no good reason, happens to intuitively determine what she has reason to do. The second stage, which involves a response to Parfit, provides a subjectivist account of non-moral cases in which subjectivism might be thought to be counterintuitive. The third stage gives an explanation of how even a subjectivist can account for moral reasons.[4]

The section then shifts more clearly to focus on the metanormative grounds of normative reasons. There is an important tradition that seeks to find space between the view that considerations are reasons as a matter of irreducibly normative fact, on the one hand, and the view that they are reasons because of some relation between the consideration and the fulfillment of our desires, on the other. According to 'constructivism,' our reasons are constructed from more basic parts of reality, including our desires or features of our rational agency. A well-known example is the Kantian approach to practical reason. In other work, Julia Markovits (2014) has defended a Kantian account that is explicitly intended to be an alternative to both Parfit's view and views like Sobel's. She had avoided calling this view 'constructivist,' however. Her joint chapter with Kenny Walden considers the prospects for a Kantian approach that

[3] See Sobel (2011, 2017) for two of his most important other works.

[4] 'Subjectivism' can also well be used to refer to the metanormative doctrine that seeks to *analyze* the normative reason-relation in terms of subjective states like desire (see, e.g., Schroeder 2007). For Sobel, however, it is a view about which considerations can be reasons. It is worth noting that both 'objectivism' and 'subjectivism' as we use them here can be both metanormative views about which considerations can be reasons and 'high-level' first-order normative views about what reasons we have – 'high level' because they would be compatible with specific substantive first-order theories like consequentialism or virtue ethics that one would more naturally discuss in normative ethics. As Berker (2018) suggests, first-order theory itself can be understood as metaphysics. Parfit's own formulation of subjectivism is ambiguous, containing the phrase "provided by, or depend upon."

is worthy of this label and develops a novel version that is informed by reflection on the problems for some existing views that go under the label.

As Markovits and Walden note, a key part of some versions of Kantian constructivism is the *constitutivist* idea that reasons derive from the *constitutive features* of agency or valuing. This constitutivist view has long been critiqued by David Enoch, who famously raised the 'Shmagency' objection to the view.[5] His chapter reviews and updates this critique, with a special focus on a version of constitutivism from Michael Smith. He argues at length that the Shmagency objection remains unanswered by Smith's view and indeed any other equally ambitious constitutivist view. At the end, he offers a small glimmer of hope for a less ambitious project: perhaps constitutivism could be scaled back so that it is combined with either a further metaphysical claim or a further normative claim to yield a package deal suited to answer the objection.

Pamela Hieronymi's chapter shows a related but different way in which a package of views about reasons and agential mental capacities might clear up some mysteries about normative metaphysics. In a series of important papers,[6] Hieronymi suggests that we can get a better understanding of the metaphysical unity of normative, motivating, and explanatory reasons by reflecting on the role that reasons play in reasoning. This chapter brings these ideas together under a unified heading – the 'Reasoning First' approach – and shows how they together avoid problems for some of the leading metaphysical accounts of reasons that the previous chapters in this subsection consider. It is for this reason that we place her chapter in this subsection near the end. But it is another example of a chapter that could be usefully read alongside those in other sections of the volume. It interacts in interesting ways with several of the chapters on motivating reasons in the first half of our third section, for example. It also exemplifies a large research program that promises to unify all three branches of the discipline. Hence, it would also be worth reading alongside the first section's chapters.

The section is rounded off by another selection from Parfit, giving the last word to non-naturalism. This selection draws from later parts of *On What Matters*, in which Parfit landed upon a new way of framing his view. The first two volumes of *On What Matters* could give one the impression that Parfit is what Enoch (2011) calls a *robust realist* about normativity, taking it to be a fundamental feature of the world. Yet Parfit had in the earlier volumes insisted that he took normative facts to exist in a 'different sense' from non-normative facts. Many were puzzled and wondered how to distinguish this view from sophisticated expressivist views that can allow it to be 'true' in an ontologically lightweight sense that there are normative facts. Parfit makes his position about the metaphysics of normativity much clearer in this selection, carving out a position he ended up calling "Non-Realist Cognitivism" and clarifying its relationship to expressivist views he had earlier opposed.

[5] See Enoch (2006, 2010) and also Railton (1997, 2004).

[6] See especially Hieronymi (2005, 2006, 2009, 2011, 2013).

b High-level substantive matters

Some questions about normative reasons are sufficiently abstract to seem unlike straightforwardly first-order questions but are also not clearly metanormative. The second subsection collects some chapters on some of these less easily classified questions.

The first two chapters consider how to understand the *force* of normative reasons. As Margaret Olivia Little and Coleen Macnamara observe at the beginning of their chapter, a surprising number of theorists assume that normative reasons are *pushy* in the following way: if a normative reason is not outweighed, one *ought* to comply with it. Some have pushed back against this pushy view over the years, arguing that some or even most normative reasons are not presumptively obliging. But there are different ways of rejecting the pushy conception that haven't been sufficiently distinguished. Little and Macnamara usefully distinguish between opponents of the pushy view who suggest that undefeated reasons generate permissions by neutralizing requirements and opponents who instead suggest that undefeated reasons are not *deontic* (i.e., suitably related to permission and obligation) but rather *commendatory*. After prying these ideas apart, they devote the chapter to explaining the different arguments for the two forms of non-requiring reason.

R. Jay Wallace's chapter comes at the topic of force from the opposite angle. In contrast to Little and Macnamara, he suggests that the recent turn toward reasons-first approaches to normativity has obscured the existence and distinctiveness of what he calls *requirements of reason*. Like reasons generally, requirements of reasons can conflict and be overridden in certain conditions. But their normative profile and function in deliberation is very different from the non-pushy reasons Little and Macnamara discuss – different enough to cast doubt on Scanlon's claim that the concept of a normative reason is just the concept of a consideration that *counts in favor* of some act or attitude. Wallace explores these points with a focus on the case of moral requirements (which he assumes to be requirements of reason), though there is some consideration of rational requirements.[7]

The next two chapters examine pluralism about practical normativity and the idea of an overall practical 'ought.' David Copp starts his chapter with a defense of a strong form of pluralism about reasons and 'oughts' according to which they are all standpoint relative (where the notion of a 'standpoint' here doesn't mean any *person's* standpoint but rather the kind we have in mind in speaking of 'the standpoint of morality'). This view may seem consistent with the thought that there is some overarching standpoint that balances the others, as has been defended by Chang (2004a, 2004b), but Copp then proceeds to argue that it isn't and hence that there is no overall practical 'ought' or overall notion of a normative practical reason: for no standpoint is neutral in the required sense.

[7] For a fuller discussion by Wallace of rational requirements, see his 2001 work.

Elizabeth Harman's chapter pairs in an interesting way with Copp's. Harman's official thesis is that there is no moral 'ought' and no prudential 'ought.' This might appear to be a straightforward denial of a claim that pluralists like Copp would want to accept. But Harman allows that there are distinctively moral *considerations* and even that there are 'distinctively moral ought facts.' She just thinks it doesn't follow that there is a distinctively moral 'ought.' Instead, the all-things-considered practical ought is the same ought in both the moral and the prudential cases and merely has a different kind of salient consideration as its normative ground in the two cases. As she puts it, her view "does not hold that there are three distinct *oughts*, one moral *ought*, one prudential *ought*, and one all-things-considered *ought*"; instead, "it is the all-things-considered *ought* that is at play throughout the phenomena we have discussed," though "[s]ome *ought* facts are *moral* facts in that they are centrally explained by moral considerations."

After this pair of chapters, the section shifts course to a chapter by Stephen Darwall that seeks to understand what makes a practical reason have peremptory force of the distinctively moral kind. Darwall's answer is informed by the *second-personal approach* he has long advocated.[8] He begins by recounting how the philosophy of practical reason after Nagel 1970 moved from a conception of practical reasons as fundamentally *first personal* (i.e., ones addressed to *me*) to a conception that took account of *third-personal reasons* (i.e., ones addressed *impersonally* to agents). He then argues that once these two perspectives are acknowledged as normatively grounding distinctive kinds of practical reason, we should expect there to be *second-personal* reasons, which are addressed from you to me or me to you. For Darwall, distinctively moral reasons are second personal. After developing this idea, Darwall uses it to explain the authority of morality.

5 Practical rationality

The final section collects some work on rationality. It is worth noting that although we placed Broome's chapter at the beginning rather than in this section, the reader would do well to reread that chapter alongside this section: Broome's work has been very influential on the branch covered by this section, and some key Broomean themes concerning this branch appear in his chapter.

The section opens with a contribution by Errol Lord that rehearses a view about practical rationality that emerged through the work of Broome, Kolodny, and Scanlon, according to which rationality is to be understood in terms of *coherence requirements*. Lord then explains why this conception of rationality seems to lead to a problem about the normativity of rationality – that is, about why it should *matter* whether we are rational or irrational. As Lord argues (see also Kolodny 2005), the best view about the

[8] The *locus classicus* is Darwall (2006).

form of coherence requirements (the 'narrow scope' view) suggests that coherence can sometimes require one to do things that one shouldn't do, in the sense of 'shouldn't' that interests us when we are deliberating. In line with his earlier work (2014, 2018), Lord suggests that we can avoid this problem and understand the normativity of rationality if we reject the view that rationality fundamentally consists of complying with requirements of coherence. We should, he argues, instead accept the view that rationality consists of responding to the balance of *possessed* normative reasons, which for him are facts that one possesses as reasons by virtue of one's being in a position to know them and how to respond to them. We can then see coherence as having derivative significance as an upshot of responding to such reasons.

Lord's chapter is followed by a chapter by one of us which is on a narrower issue about the normativity of rationality. As Sylvan begins by noting, one of the first papers to raise a problem about the normativity of rationality was Raz's (2005) 'The Myth of Instrumental Rationality.' While this paper officially had a much narrower focus than the influential piece by Kolodny published in the same year, Raz's challenges turned out to be special cases of the broader ones that Lord discusses. Sylvan reopens Raz's question by asking whether there might after all be a *special* problem about instrumental rationality. Sylvan thinks there is. He gives five new arguments for skepticism about instrumental rationality, some of which are inspired by Continental figures who offered critiques of instrumental reason (e.g., Arendt, Horkheimer, Weber, and André Gorz). After concluding that these arguments support skepticism about the normativity of instrumental rationality, Sylvan suggests that we can capture the phenomena that instrumental principles were meant to capture with certain non-instrumental coherence requirements and thereby avoid the special problems. His way of avoiding the problems is compatible with Lord's thought that rationality is not *just* coherence. But it is also compatible with a mixed view which sees rationality as essentially including some coherence requirements, including the non-instrumental ones he defends. This story fits nicely, as he notes at the end, with a story that he has given about the non-coherentist sides of practical and epistemic rationality in Sylvan (forthcoming, 2020).

The chapters by Lord and Sylvan are followed by three chapters on a different side of rationality in the narrow sense. In the literature on rationality that emerged from Broome's work, it has been common to focus on *synchronic* rational requirements. But rational requirements can be construed as *process requirements*, compliance with which unfolds over time. Kolodny (2005) emphasized this point in defending the 'narrow scope' account of rational requirements. But well before that, two rich literatures emerging from decision theory, on the one hand, and the work of Michael Bratman (e.g., Bratman 1987), on the other, were independently guided by the idea that practical coherence is partly a diachronic matter. The chapter by Chrisoula Andreou, Bratman, and Wlodek Rabinowicz come at practical rationality from these angles. Bratman examines diachronic rationality in light of his planning theory of agency, Rabinowicz looks at how to account for diachronic requirements from a decision-theoretic perspec-

tive, and Andreou explores diachronic requirements while drawing on knowledge of both the decision-theoretic and the Bratman-inspired literatures.

Building on her earlier work on the topic,[9] Andreou looks at three kinds of cases which seem to involve diachronic irrationality: cases of preference reversal due to temptation, cases in which vague goals lead to procrastination, and cases in which deliberation is delayed by cycling between incommensurable (or incomparable) objects of choice. She is careful throughout to distinguish structural criticisms of such agents (e.g., their preferences violate an alleged transitivity requirement) and more substantive criticisms (e.g., they don't sufficiently value their well-being over time or have poor managerial skills). She considers several structural diagnoses of what is going wrong in these cases, aiming more to introduce the reader to the terrain than to defend a particular diagnosis. The chapter does an outstanding job providing a balanced, unified view of the decision-theoretic and philosophical literatures.

Bratman's chapter begins by rehearsing the challenge to the normativity of synchronic coherence requirements and his earlier effort (see Bratman 2009b) to address this challenge by appealing to a reason for *self-governance*. He then explains how to extend his self-governance approach to synchronic coherence requirements on planning at a time to explain diachronic practical coherence, including some of cases that interested Andreou (e.g., being led off course by temptation and cycling through options). The result is a unified self-governance–based account of the normativity of both synchronic and diachronic instrumental rationality. His appeal to self-governance is part of a larger package of views that includes a broadly pragmatic 'two-tiered' defense of the rationality of norms, a defense that is sensitive to human limitations.

Rabinowicz's chapter addresses diachronic rationality from a decision-theoretic perspective. He focuses on how decision theory should advise agents who fail to stick to their plans owing to a failure to be expected utility-maximizers and in particular who violate the Independence Axiom of expected utility theory. He rehearses two standard approaches to this question, which recommend policies of 'sophisticated choice' and 'resolute choice' (in some specific technical senses of those phrases) and then discusses and further develops a 'wise choice' policy which synthesizes these approaches, which he defended earlier in Rabinowicz (1995). Like sophisticated choice, wise choice makes use of backward induction in reasoning about sequential decision problems, but like resolute choice, it rejects pure future-directedness of sophisticated choice and makes room for commitments to previously adopted plans. The chapter also considers whether wise choice can be reduced to sophisticated choice as an appropriate re-description of the decision problem. Although this chapter is the most formal of all, it is written to be accessible to outsiders; close study of it alongside a big-picture work like Buchak (2013) will introduce the reader to formal philosophy of practical reason and showcase a novel view in it.

[9] See, for example, Andreou (2014, 2015, 2016).

The book then ends with a chapter by Jennifer Morton and Sarah Paul which provides an important and fresh contrast to the Broomean approach to this branch, and which also ties together all three branches of the philosophy of practical reason. Like Broome (2013), Morton and Paul see an important connection between rationality and *norms of reasoning*. But they argue that this connection favors a very different account of rationality than Broome's. For Broome, the rational requirements which underpin good reasoning are *a priori* and categorical. Morton and Paul reject these ideas and defend what they call an *ecological* approach, which draws on insights from the tradition of bounded rationality.[10] In this approach, the norms that underwrite good reasoning for a given agent should be sensitive to the distinctive features of that agent's circumstances and psychology and hence are not *a priori* or purely structural. Their approach provides a distinctive vindication of the normativity of rationality. While we placed it in this section, it would also be worth reading alongside the chapters in the first and third sections and illustrates the empirically informed approach also displayed by Anderson, Haslanger, Tiberius and Washington, and Wonderly.

* * *

The literature on practical reason is a quicksand with unclear boundaries. It would take a series of handbooks to survey it as completely as we would like. But our job was to make one book. Inevitably, then, there are unexamined angles on topics that are covered, topics that are not covered, and adjacent literatures that some might classify as part of the philosophy of practical reason but which we have mostly bracketed. We hope, however, that the chapters in this volume provide readers with an overview of the main issues in this fast-evolving field, along with a sense of its richness and depth. We invite readers to peruse the 'Guide to Further Reading,' given as an appendix to this volume, should they wish to continue their investigations.

[10] This tradition is rooted in the work of economist and cognitive psychologist Herbert Simon (e.g., 1955).

References

Andreou, C. 2014. 'Temptation, Resolutions, and Regret.' *Inquiry* 57: 275–292.

Andreou, C. 2015. 'The Real Puzzle of the Self-Torturer: Uncovering a New Dimension of Instrumental Rationality.' *Canadian Journal of Philosophy* 45: 562–575.

Andreou, C. 2016. 'Dynamic Choice' in *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/dynamic-choice/.

Anscombe, G. E. M. 1957. *Intention*. Oxford: Blackwell.

Arpaly, N. 2000. 'On Acting Rationally Against One's Best Judgment.' *Ethics* 110: 488–513.

Arpaly, N. 2003. *Unprincipled Virtue*. Oxford: Oxford University Press.

Arpaly, N. and Schroeder, T. 2012. 'Deliberation and Acting for Reasons.' *Philosophical Review* 121: 209–239.

Berker, S. 2018. 'The Unity of Grounding.' *Mind* 127: 729–777.

Brandt, R. 1979. *A Theory of the Good and the Right*. Oxford: Clarendon Press.

Bratman, M. 1987. *Intention, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press.

Bratman, M. 1999. *Faces of Intention*. Cambridge: Cambridge University Press.

Bratman, M. 2009a. 'Intention, Belief, and Practical Rationality' in Sobel, D. and Wall, S. (eds.) *Reasons for Action*. Cambridge: Cambridge University Press.

Bratman, M. 2009b. 'Intention, Practical Rationality, and Self-Governance.' *Ethics* 119: 411–443.

Bratman, M. 2018. Planning, Time, and Self-Governance: Essays in Practical Rationality. *Oxford: Oxford University Press.*

Broome, J. 1999. 'Normative Requirements.' *Ratio* 12: 398–419.

Broome, J. 2004. 'Reasons' in Wallace, R. J., Smith, M., Scheffler, S. and Pettit, P. (eds.) *Reason and Value: Themes from the Philosophy of Joseph Raz*. Oxford: Oxford University Press.

Broome, J. 2005. 'Does Rationality Give Us Reasons?' *Philosophical Issues* 15: 321–337.

Broome, J. 2007a. 'Does Rationality Consist in Correctly Responding to Reasons?' *Journal of Moral Philosophy* 4: 349–374.

Broome, J. 2007b. 'Wide or Narrow Scope?' *Mind* 116: 359–370.

Broome, J. 2013. *Rationality through Reasoning*. Oxford: Blackwell.

Buchak, L. 2013. *Risk and Rationality*. Oxford: Oxford University Press.

Callard, A. 2018. *Aspiration*. Oxford: Oxford University Press.
Chang, R. (ed.) 1997. *Incommensurability, Incomparability, and Practical Reason*. Cambridge, MA: Harvard University Press.
Chang, R. 2004a. 'Putting Together Morality and Well-Being' in Betzler, M. and Baumann, P. (eds.) *Practical Conflicts*. Cambridge: Cambridge University Press.
Chang, R. 2004b. 'All Things Considered.' *Philosophical Perspectives* 18: 1–22.
Chang, R. 2009. 'Voluntarist Reasons and the Sources of Normativity' in Sobel, D. and Wall, S. (eds.) *Reasons for Action*. Cambridge: Cambridge University Press.
Chang, R. 2013a. 'Commitment, Reasons, and the Will' in Shafer-Landau, R. (ed.) *Oxford Studies in Metaethics*, 8. Oxford: Oxford University Press.
Chang, R. 2013b. 'Grounding Practical Normativity: Going Hybrid.' *Philosophical Studies* 164: 163–187.
Chang, R. 2015. 'Comparativism: The Ground of Rational Choice,' in Errol Lord and Barry McGuire (eds.) *Weighing Reasons*. Oxford: Oxford University Press.
Chang, R. 2017. 'Hard Choices.' *APA Journal of Philosophy* 92: 586–620.
Dancy, J. 1977. 'The Logical Conscience.' *Analysis* 37: 81–84.
Dancy, J. 1993. *Moral Reasons*. Oxford: Blackwell.
Dancy, J. 2000. *Practical Reality*. Oxford: Oxford University Press.
Dancy, J. 2004. *Ethics without Principles*. Oxford: Oxford University Press.
Dancy, J. 2006. 'Nonnaturalism' in Copp, D. (ed.) *The Oxford Handbook of Ethical Theory*. Oxford: Oxford University Press.
Dancy, J. 2018. *Practical Shape*. Oxford: Oxford University Press.
Darwall, S. 1983. *Impartial Reason*. Ithaca: Cornell University Press.
Darwall, S. 2006. *The Second-Person Standpoint*. Cambridge, MA: Harvard University Press.
Davidson, D. 1963. 'Actions, Reasons, and Causes.' *Journal of Philosophy* 60: 685–700.
Davidson, D. 1970. 'Mental Events' in Foster, L. and Swanson, J. W. (eds.) *Experience and Theory*. London: Duckworth.
Dretske, F. 1988. *Explaining Behavior*. Cambridge, MA: The MIT Press.
Enoch, D. 2006. 'Agency, Shmagency: Why Normativity Won't Come from What Is Constitutive of Agency.' *Philosophical Review* 115: 169–198.
Enoch, D. 2010. 'Shmagency Revisited' in Brady, M. (ed.) *New Waves in Metaethics*. Basingstoke: Palgrave.
Enoch, D. 2011. Taking Morality Seriously: A Defense of Robust Realism. *Oxford: Oxford University Press.*
Fix, J. D. 2018. 'Intellectual Isolation.' *Mind* 127: 491–520.
Foot, P. 1972a. 'Morality as a System of Hypothetical Imperatives.' *Philosophical Review* 81: 305–316.
Foot, P. 1972b. 'Reasons for Action and Desires.' *Proceedings of the Aristotelian Society Supplementary Volume* 46: 203–210.
Foot, P. 1978. *Virtues and Vices*. Oxford: Oxford University Press.
Foot, P. 2001. *Natural Goodness*. Oxford: Oxford University Press.

Gert, J. 2007. 'Normative Strength and the Balance of Reasons.' *Philosophical Review* 116: 533–562.
Greenspan, P. 1975. 'Conditional Oughts and Hypothetical Imperatives.' *Journal of Philosophy* 72: 259–276.
Greenspan, P. 1988. *Emotions and Reasons*. London: Routledge.
Hieronymi, P. 2005. 'The Wrong Kind of Reason.' *Journal of Philosophy* 102: 437–457.
Hieronymi, P. 2006. 'Controlling Attitudes.' *Pacific Philosophical Quarterly* 87: 45–74.
Hieronymi, P. 2009. 'Two Kinds of Agency' in O'Brien, L. and Soteriou, M. (eds.) *Mental Actions*. Oxford: Oxford University Press.
Hieronymi, P. 2011. 'Reasons for Action.' *Proceedings of the Aristotelian Society* 111: 407–427.
Hieronymi, P. 2013. 'The Use of Reasons in Thought (and the Use of Earmarks in Arguments).' *Ethics* 124: 114–127.
Hill, T. 1973. 'The Hypothetical Imperative.' *Philosophical Review* 82: 429–450.
Kiesewetter, B. 2017. *The Normativity of Rationality*. Oxford: Oxford University Press.
Kolodny, N. 2005. 'Why Be Rational?' *Mind* 114: 509–563.
Kolodny, N. 2007. 'How Does Coherence Matter?' *Proceedings of the Aristotelian Society* 107: 229–263.
Korsgaard, C. 1986. 'Skepticism about Practical Reason.' *Journal of Philosophy* 83: 5–25.
Korsgaard, C. 1996a. *Creating the Kingdom of Ends*. Cambridge: Cambridge University Press.
Korsgaard, C. 1996b. *The Sources of Normativity*. Cambridge: Cambridge University Press.
Korsgaard, C. 1997. 'The Normativity of Instrumental Reason' in Cullity, G. and Gaut, B. (eds.) *Ethics and Practical Reason*. Oxford: Clarendon Press.
Korsgaard, C. 2008. *The Constitution of Agency*. Oxford: Oxford University Press.
Korsgaard, C. 2009a. *Self-Constitution*. Oxford: Oxford University Press.
Korsgaard, C. 2009b. 'The Activity of Reason.' Proceedings and Addresses of the American Philosophical Association *83: 23–43.*
Lord, E. 2014. 'The Coherent and the Rational.' *Analytic Philosophy* 55: 151–175.
Lord, E. 2018. *The Importance of Being Rational*. Oxford: Oxford University Press.
Lord, E. and Maguire, B. (eds.) 2016. *Weighing Reasons*. Oxford: Oxford University Press.
Mackie, J. L. 1977. *Ethics: Inventing Right and Wrong*. London: Pelican Books.
Manne, K. 2014. 'Internalism about Reasons: Sad but True?' *Philosophical Studies* 167: 89–117.
Markovits, J. 2014. *Moral Reason*. Oxford: Oxford University Press.
McDowell, J. 1978. 'Are Moral Requirements Hypothetical Imperatives?' *Proceedings of the Aristotelian Society Supplementary Volume* 52: 13–29.
McDowell, J. 1979. 'Virtue and Reason.' *The Monist* 62: 331–350.

McDowell, J. 1995. 'Might There Be External Reasons?' in Altham, J. E. J. and Harrison, R. (eds.) *World, Mind, and Ethics: Essays on the Ethical Philosophy of Bernard Williams*. Cambridge: Cambridge University Press.

McDowell, J. 1998. *Mind, Value and Reality*. Cambridge, MA: Harvard University Press.

Mele, A. 1992. *Springs of Action*. Oxford: Oxford University Press.

Mele, A. 2003. *Motivation and Agency*. Oxford: Oxford University Press.

Mele, A. and Rawling, P. (eds.) 2004. *The Oxford Handbook of Rationality*. Oxford: Oxford University Press.

Millgram, E. (ed.) 2001. *Varieties of Practical Reasoning*. Cambridge, MA: The MIT Press.

Nagel, T. 1970. *The Possibility of Altruism*. Princeton: Princeton University Press.

Nagel, T. 1986. *The View from Nowhere*. Oxford: Oxford University Press.

Parfit, D. 1997. 'Reasons and Motivation.' *Aristotelian Society Supplementary Volume* 70: 99–130.

Parfit, D. 2001. 'Rationality and Reasons' in Egonsson, D., Josefsson, J., Petersson, B., and Ronnow-Rasmussen, T. (eds.) *Exploring Practical Philosophy*. Burlington, VT: Ashgate.

Parfit, D. 2011. *On What Matters, Volumes 1 and 2*. Oxford: Oxford University Press.

Parfit, D. 2017. *On What Matters, Volume 3*. Oxford: Oxford University Press.

Paul, L. A. 2014. *Transformative Experience*. Oxford: Oxford University Press.

Rabinowicz, W. 1995. 'To Have One's Cake and Eat It, Too: Sequential Choices and Expected-Utility Violations.' *The Journal of Philosophy* 92: 586–620.

Railton, P. 1986. 'Moral Realism.' *Philosophical Review* 95: 163–207.

Railton, P. 1997. 'On the Hypothetical and Non-Hypothetical in Reasoning about Belief and Action' in G. Cullity and B. Gaut, (eds.) *Ethics and Practical Reason.* Oxford: Oxford University Press.

Railton, P. 2004. 'How to Engage Reason: The Problem of Regress' in Wallace, J. et al. (eds.) *Reason and Value: Themes from the Moral Philosophy of Joseph Raz*. Oxford: Oxford University Press.

Raz, J. 2002. *Engaging Reason*. Oxford: Oxford University Press.

Raz, J. 2005. 'The Myth of Instrumental Rationality.' *Journal of Ethics and Social Philosophy* 1: 1–28.

Raz, J. 2011. *From Normativity to Responsibility*. Oxford: Oxford University Press.

Richardson, H. 1994. *Practical Reasoning about Final Ends*. Cambridge: Cambridge University Press.

Rödl, S. 2007. *Self-Consciousness*. Cambridge, MA: Harvard University Press.

Rosati, C. 1996. 'Internalism and the Good for a Person.' *Ethics* 106: 297–326.

Rosati, C. 2016. 'Moral Motivation' in *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/moral-motivation/.

Scanlon, T. M. 1998. *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.

Scanlon, T. M. 2007. 'Structural Irrationality' in Brennan, G., Goodin, R., Jackson, F. and Smith, M. (eds.) *Common Minds: Themes from the Philosophy of Philip Pettit*. Oxford: Clarendon Press.
Scanlon, T. M. 2013. *Being Realistic about Reasons*. Oxford: Oxford University Press.
Schapiro, T. 2001. 'Three Conceptions of Action in Moral Theory.' *Nous* 35: 93–117.
Schroeder, M. 2007. *Slaves of the Passions*. Oxford: Oxford University Press.
Schroeder, M. 2011. 'Holism, Weight and Undercutting.' *Nous* 45: 328–344.
Setiya, K. 2007. 'Cognitivism about Instrumental Reason.' *Ethics* 117: 649–673.
Simon, H. 1955. 'A Behavioral Model of Rational Choice.' *Quarterly Journal of Economics* 69: 99–118.
Sinhababu, N. 2017. Humean Nature: How Desire Explains Action, Thought, and Feeling. *Oxford: Oxford University Press.*
Skorupski, J. 2010. *The Domain of Reasons*. Oxford: Oxford University Press.
Skorupski, J. 2017. 'Reply to Sylvan: Constructivism? Not Kant, Not I.' *Philosophical Quarterly* 67: 593–605.
Smith, M. 1994. *The Moral Problem*. Oxford: Blackwell.
Smith, M. 1995. 'Internal Reasons.' *Philosophy and Phenomenological Research* 55: 109–131.
Smith, M. 2004. *Ethics and the A Priori*. Cambridge: Cambridge University Press.
Sobel, D. 2011. 'Parfit's Case against Subjectivism.' *Oxford Studies in Metaethics* 6: 52–78.
Sobel, D. 2017. *From Valuing to Value*. Oxford: Oxford University Press.
Star, D. (ed.) 2018. *The Oxford Handbook of Reasons and Normativity*. Oxford: Oxford University Press.
Street, S. 2008. 'Constructivism about Reasons.' *Oxford Studies in Metaethics* 3: 207–245.
Street, S. 2009. 'In Defense of Future Tuesday Indifference: Ideally Coherent Eccentrics and the Contingency of What Matters.' *Philosophical Issues*, 273–298.
Sylvan, K. 2020. 'An Epistemic Non-Consequentialism.' *Philosophical Review* 129: 1–51.
Sylvan, K. Forthcoming. 'Respect and the Reality of Apparent Reasons.' *Philosophical Studies*.
Tenenbaum, S. 2007. Appearances of the Good: An Essay on the Nature of Practical Reason. *Cambridge: Cambridge University Press.*
Thompson, M. 2008. *Life and Action*. Cambridge, MA: Harvard University Press.
Vogler, C. 2002. *Reasonably Vicious*. Cambridge, MA: Harvard University Press.
Wallace, R. J. 2001. 'Normativity, Commitment, and Instrumental Reason.' *Philosophers' Imprint* 1: 1–26.
Way, J. 2010. 'Defending the Wide Scope Approach to Instrumental Reason.' *Philosophical Studies* 147: 213–233.
Williams, B. 1979. 'Internal and External Reasons' reprinted in *Moral Luck* (1980). Cambridge: Cambridge University Press.

Part 1: Foundational matters

1. Some central questions about practical reason

1 Introduction

The main aim of this chapter is to call attention to some questions that I believe students and others who are interested in practical rationality should attend to. I of course have my own views about the best answers to these questions, and I will indicate what these answers are. But my main aim will be just to identify the questions that seem to me important and discuss some things that need to be considered in answering them.

I am a cognitivist about normative judgments. I believe that normative judgments are capable of being true or false and that accepting such a judgment is a form of belief. The realm of normative judgments, as I understand it, includes not only moral judgments about right and wrong but also judgments about reasons for action and for beliefs and other attitudes, judgments about what individuals ought to do, and judgments about what is good. Since I believe that normative judgments of other kinds are best understood in terms of reasons, I will focus on judgments about reasons. But the main points I will make do not depend on the view that reasons are fundamental in this way. The questions I will call attention to, and my responses to them, could be stated as well in terms of other normative relations.

Some writers who, like me, are cognitivists about normative judgments distinguish their view from realism about normative facts – the ontological thesis that such facts exist – which they see as having potentially troubling ontological implications. (For example, Skorupski 2010; Parfit 2011.) Since I believe that if normative judgments can be true, then normative facts exist in the only sense of existence that is relevant to them (Scanlon 2014; Scanlon 2017), I am a normative realist as well as a cognitivist. As I will indicate in the next section, I do not believe that the existence of normative facts has implications that we should find troubling. But, as I will also say, whether this is so is one of the questions about practical reason that needs to be addressed.

2 Common objections to normative realism

Realism about normative judgments has been seen as subject to three seemingly strong objections, famously stated by John Mackie: metaphysical "queerness," motivational impotence, and epistemological obscurity. (Mackie 1977) The first questions I will call attention to concern not only how to respond to these objections but also, more deeply, how the objections themselves should be understood and why they should be seen as challenging.

2.1 Metaphysics

Mackie famously wrote that facts about moral rightness and wrongness, insofar as they are understood to involve objective prescriptivity, would be metaphysically queer, "utterly different from anything else in the universe" (Mackie 1977: 38).[1] Contemporary anti-realists say the same about facts about reasons for action, often putting this as the claim that such facts would be incompatible with a scientific view of the world. The questions that need to be addressed here are:

> *(1)* Is there a problem about how normative facts could be a part of the world? What "world" is in question?

If the world in question were the natural world of physical objects, causes, and effects (as the word 'universe' might suggest), then this objection would have force. Normative facts would be unlike anything else in this world. But realism about normative truth of the kind I am defending does not claim that normative facts and properties are parts of the natural world. Those of us who defend this view are *non-naturalists*. We are explicitly *not* claiming that normative truths state facts about the natural world, and it is the non-naturalist character of the view that those who raise this objection find implausible.

[1] I assume that Mackie would also have objected to objective truths about reasons for action, although it is not clear that he would have been consistent in doing this, insofar as he was committed to there being truths about the reasons that individuals have, given their desires. Although he no doubt would have referred to these reasons as subjective, the claim that individuals who have the certain desires have reason to do what would promote them would seem to be an objective claim.

So what world, or universe, is it that (1) normative facts would have to be part of in order for there to be normative truths,[2] and (2) it is metaphysically implausible that this world should include such facts? Perhaps it is just the world of things that we are ontologically committed to. But in order for ontological commitment to certain things, such as abstract entities or normative facts, to be implausible, the idea of existence that ontological commitment commits one to has to have some content.

My view is that claims of existence that have content are all "domain specific." That is to say, the content of such claims depends upon the subject matter in question. Existence of physical objects is one thing (a matter of having such things as spatio-temporal location, causal interaction, and so on.) Existence of numbers and sets is something different (a purely mathematical matter.) And the existence of normative facts and relations is something else altogether. There is no broader "world" which all of these things are part of insofar as they exist. In particular, physical objects and natural properties do not exist in some broader sense of this kind, in addition to being parts of the physical world. They exist only in a domain-specific sense. I have argued for this view at length elsewhere (Scanlon 2014: 16–30; Scanlon 2017). My main point here is the broader one that in order to assess the metaphysical objection to normative realism, one needs to be clear about what world, or idea of existence, is in question.

2.2 Motivation

Turning now to the motivational objection, the problem is supposed to be that a cognitivist view, according to which accepting a normative judgment is a matter of having a certain belief, would be unable to explain the connection between accepting such judgments and acting in certain ways. The questions I want to call attention to here are:

> *(2)* Is there a problem about how a cognitivist view of normative judgments could explain the connection between these judgments and actions? What is the connection that needs to be explained?

The term 'motivation' suggests that the connection in question is a psychological one about how the presence of a belief about reasons could causally explain subsequent action in accord with the normative content of that belief. But the connection is not only causal. Even Donald Davidson, in his classic statement of the view of reasons as causes (1980), said that reasons not only cause actions, they also "rationalize them." And the term 'motivation' itself has a "rationalizing" aspect. The question of what motivated an agent to do a certain thing is a question about what reason she saw for doing it, not only (or even, I would say, primarily) about what caused her to act in that way. The importance of this rational aspect of the connection between normative

[2] A question well-raised in (McDowell 1985) and (Tait 2005: 8).

attitudes and action is evident also from what non-cognitivists say about the matter. R. M. Hare, for example, said that moral judgments had to involve the acceptance of an imperative because imperatives were the only kinds of utterance that were logically tied with action (1952: 20, 171–172). By this I think he meant that only an interpretation of such judgments as involving imperatives can explain the fact that the acceptance of a normative judgment can make it rational for a person to act in a certain way and even irrational not to so act.

This suggests a "two-track" account of the connection between normative judgments and action, based on the idea of a rational agent. On the one hand, as Hare seems to be saying, it is irrational not to act in accord with imperatives that one sincerely accepts. This captures the "rationalizing" aspect of the idea of motivation. But, on the other hand, rational agents (at least of the embodied kind that we are familiar with) are so constituted, physically, that they normally act in accord with the imperatives they sincerely accept. This is not to say that imperatives (or mental states of accepting imperatives) are causes but only that there is some causal story that explains the uniformities in behavior typical of a rational agent.

A two-track explanation of this kind is equally available to a cognitivist. A rational agent is a being that is capable of arriving at judgments about the reasons it has and is irrational if it fails to act in accord with these judgments. Moreover, rational agents are so constituted physically that normally, although not invariably, they act in accord with these judgments. The normative judgments that such a being accepts thus rationalize certain actions (make it rational for the being to act in certain ways and irrational for it not to do so.) And these uniformities are underwritten by some causal mechanism. An explanation of the connection between normative judgment and action along these lines seems to me extremely plausible (Scanlon 2014: 54–58).

2.3 Epistemology

Any account of normative truth needs to be compatible with some explanation of how our normative beliefs can depend on, and be responsive to, the normative facts. It would thus be a serious objection to normative realism if, as Mackie and others have maintained, it ruled out any explanation of this kind. So we need answers to the following questions:

> *(3)* Is there a problem about how we could come to know normative truths if normative realism were correct? What is the problem, and what would a plausible epistemology of normative belief have to be like?

There would be no problem of this kind for an account according to which normative facts are, at the most basic level, dependent on our beliefs about them or on our other attitudes. There would be no such problem, for example, for a reductive desire theory, according to which facts about reasons for action just are facts about which actions

will promote the satisfaction of our desires (Schroeder 2007). Assuming that we have access to our own desires, and can form reliable beliefs about causes and effects in the natural world, it would be not at all mysterious how we could arrive at true normative beliefs on this account.

But if the normative facts are independent of us, there may seem to be a problem about how our beliefs could be responsive to these facts. This might not be a problem if normative facts or properties had causal powers. But non-naturalists deny that this is so. Paul Benacerraf (1973) famously argued that the fact that we have no causal interaction with mathematical facts or entities represented a serious problem for realist interpretations of mathematical truth, and his argument may seem to apply to beliefs about other abstract domains, including the normative domain. But causal interaction is not the only plausible explanation of belief formation, and it seems particularly unsuited to the case of abstract beliefs.[3]

So some alternative explanation is needed, one that is more plausible than the idea that we can be aware of the facts about a domain through a special faculty of "intuition" that is a non-causal analog of perception. The most plausible response to this problem seems to me to lie along the following lines. We arrive at beliefs about abstract subjects by reasoning about them in the right way. Two things are required in order to explain, for a given subject matter, what this involves and how it is possible. First, we need an understanding of that subject that provides the basis for a clear idea of what "reasoning about it in the right way" involves. Second, it must be plausible to believe that we are capable of engaging in that kind of thinking.

In the case of arithmetic, for example, an understanding of the natural numbers provides the basis for explaining why counting, arithmetical calculation, and reasoning by mathematical induction are "right ways" of forming arithmetical beliefs. Given such an account, the dependence of our beliefs about a subject on the facts about it can be explained by the fact that we have the capacity to engage in the relevant forms of reasoning. Thinking about "the number line" involves a kind of mental picturing, but this is not properly understood on the model of perception, and there is nothing mysterious about it.

Here there is a sharp difference between normative truth and mathematical truth. In the case of arithmetical truth, and to an extent truth about set theory, we have an overall conception of the subject, in mathematical terms, which provides the basis for at least a provisional account of the kind of reasoning that is involved in discovering the truth about those subjects. In the case of normative beliefs, however, we do not have a comparably systematic account of the normative domain, which can provide the basis for an account of what good normative reasoning amounts to. It is this substantive incompleteness (at least in our understanding) of the normative domain, rather than

[3] For critical discussion of how Benacerraf's argument should be understood and whether it really raises a problem for realist interpretations of mathematical truth or normative truth, see (Clarke-Doane 2017; Tait 1986).

an epistemological difficulty about how we could "be in touch with" normative facts, that presents a problem for a realist view of normative truth.

It has also been questioned whether, given our evolutionary history, we have the capacity to engage in the kind of reasoning about normative truth that a realist view would require. There is no problem of this kind in the case of arithmetic, since it seems clear that the ability to count and to reason about arithmetical relations would have been an important evolutionary advantage for our distant ancestors. It has been argued, however, that the capacity to discern the normative truth, on a realistic construal, would convey no such advantage. Given that our current evaluative attitudes are in large part results of our evolutionary history, there is no reason to believe that these attitudes tend to track the normative truth, as realists understand it. Thus, it is claimed, there is good reason to doubt that we actually have the ability to engage in normative thinking of kind that normative realism would require.[4] Assessing this challenge is thus one part of answering question (3).

[4] This version of the challenge is developed by Sharon Street (Street 2006; Street 2013) For other versions, see (Joyce 2001; Bedke 2009). For a critical response to Street's argument, see (Berker 2017).

3 Reasons and rationality

I turn now to a set of questions about the relation between rationality and substantive practical truths about what to do or think. The first of these questions is:

> *(4)* Can facts about what an individual has reason to do be based on an idea of rationality?

There are several reasons for wanting to base reasons on rationality in this way. First, an account of normative truths that based them in an idea of rationality might offer answers to the three objections I mentioned at the outset, having to do with metaphysics, motivation, and epistemology. Second, basing truths about reasons on an idea of rationality might provide a systematic overall account of normative truths of the kind that I have mentioned, a desirable alternative to a view of the normative domain as a collection of isolated facts about particular reasons and their strengths.

Third, grounding truths about reasons in an idea of rationality could explain what might be called the authority of reasons for the person for whom they are reasons. If facts about reasons were simply facts "about the world," independent of the agent, then one might ask why they are something the person should recognize as things to be guided by. One might ask, as Christine Korsgaard does, how facts about reasons "get a grip" on an agent (1996: 44–46). My own view is that the authority of truths about reasons is purely normative: it lies just in the normative fact of *being a reason for* an agent in his or her circumstances (Scanlon 2014: 10). But many others believe that some further explanation is required.

3.1 Three ideas of rationality

One way of providing such an explanation would be to show that facts about the reasons a person has are things that the person must recognize as action guiding insofar as he or she is rational. To assess the prospects for an account of this kind, it is important to distinguish three distinct ideas of rationality that might be appealed to: the idea of a rational being, the idea of the rational thing to do (what a person has most reason to do), and requirements of rationality that a person cannot violate without being *irrational*. Let me say something about each of these ideas and explain why the differences between them are important.

As I said earlier, a rational being is one that has the capacity to think about what reasons it has and to decide what to do, what to believe, and what other attitudes to

adopt, in a way that is responsive to these reasons. Rational beings need not always do this perfectly. They can have mistaken beliefs about the reasons they have, and they can even fail to do what they themselves believe is supported by the reasons they take themselves to have.

A different idea that is often called rationality is the idea of what a person has most reason to do, or would do if he or she were "perfectly rational" – that is to say, if she were not mistaken about the reasons she has given the non-normative facts available to him or her and perfectly responsive to the reasons she takes herself to have. For example, one "conception of rationality" in this sense is that what it is rational for people to do is always to act in their self-interest. I do not think that this claim about the reasons people have is correct. Whether it is correct or not, however, I do not think it is helpful to see it as a claim about *rationality*. To do so simply builds into the idea of rationality a particular view about what reasons we have that is not connected to or grounded in the idea of rationality itself.

A third idea of rationality is the idea of the requirements that an agent must satisfy in order not to be *irrational*. This idea is sometimes not distinguished from the previous one, as when, for example, it is said that it is "irrational" to knowingly act contrary to one's self interest. But I believe that this is a mistake: not every case of acting on mistaken views about the reasons one has should be counted as irrational. Nor, I would say, is it irrational to hold contradictory beliefs if one does not realize that this is so.

One thing that *is* irrational is to fail act, or to form one's beliefs or other attitudes, in a way that is supported by the reasons one believes oneself to have. It is irrational, for example, to continue to believe something that one believes there is conclusive evidence against or to do something that one believes one has conclusive reason not to do. These are violations of what John Broome calls the *enkratic* condition, which requires acting in accord with the reasons one believes oneself to have. (2007, 2013) But not all instances of irrationality in the sense I am now discussing (what I have called cases of structural irrationality [Scanlon 2007]) are violations of this requirement. Failures of instrumental reasoning, for example, need not be such violations. If I believe that I have strong reason to be in Chicago by tomorrow morning, then it is irrational of me to deny that the fact that I need to leave for the airport now in order to get to Chicago by tomorrow morning is a reason to leave for the airport now.

These three ideas of rationality differ in their suitability for the strategy of grounding claims about reasons in claims about rationality. An idea of rationality in the second sense just discussed – a conception of what individuals have most reason to do – would not be suitable for this strategy, since it would just build in at the start the conclusions about reasons that one was supposed to be grounding. So the strategy needs to employ an idea of rationality in some other sense.

The third sense of rationality described previously is better suited for this role. It supports conclusions about reasons that a person cannot consistently reject, given the other attitudes that he or she has. Moreover, requirements of rationality in this sense

are independent of particular claims about the reasons people have.[1] In the case of the person who has the aim of getting to Chicago by morning, for example, even if he has no reason at all to go to Chicago, as long as he has this aim, this conception of rationality requires him to see himself as having reason to take relevant means.

3.2 The normativity of requirements of (structural) rationality

This independence of the reasons that a person has raises a question about these requirements of rationality:

(5) In what sense are requirements of (structural) rationality normative?

When we say that the person in my example "must" see the fact that he needs to leave for the airport now in order to get to Chicago by morning as a reason to leave for the airport, what is the nature of this "must?"

One possible response would appeal to the first sense of rationality that I mentioned previously, the idea of a rational agent. A rational agent will tend to act on the reasons she judges herself to have and will see herself as having reason to take means to those ends that she has and takes herself to have sufficient reason to have. A perfectly rational agent will always do these things. So requirements of structural rationality are standards of proper functioning as a rational agent.

The question I want to call attention to, however, is whether this idea of functioning well as a rational agent plays any normative role for the agent him- or herself. If we were to ask the person in my example why he was calling for a taxi. It seems unlikely he would say that otherwise he would not be functioning properly as a rational agent. More likely, he would cite his need, or desire, to get to Chicago and, ultimately, his reason for wanting to be there. If he saw no reason to get to Chicago by morning, it would be irrational of him to have this aim and difficult to imagine his seeing any reason to do what he takes to be necessary to get there. So, although the normative force of instrumental reasoning in this case is, as I said, independent of whether he actually has any reason to get to Chicago by morning, its force, for him, does not seem to be independent of whether he takes himself to have such a reason.

This suggests that the apparent normative force of requirements of structural rationality for agents themselves is provided by the particular reasons they take themselves to have for the aims in question rather than from any separate normativity of the

[1] As Broome says, whether a person is irrational in this sense depends only on the relations between that person's own attitudes, not on the correctness of those attitudes (2007). Derek Parfit, by contrast, argues that it can be irrational to deny some truths about reasons if these are sufficiently obvious (2011 Vol. I: 120–124).

requirements themselves. Views of this kind have been put forward in different forms by Kolodny and Raz (Kolodny 2005; Raz 2005; Raz 2009).

Even if the reasons a person sees for adopting an aim play a crucial role in explaining the rationality of taking means to promote that aim, the fact that the person has adopted the aim also makes a difference to what she must do insofar as she is not irrational and even to what reasons she has. If I have decided (for good reason) to go to Chicago in December, then I have reason to pack a warm coat or buy or borrow one if I do not already have one. I would not have this reason if I had decided instead to go to Los Angeles, which I had equally good reasons to do. It is a very interesting question, on which there is a wide literature, how this is best explained.[2]

3.3 Anchoring reasons in an agent's attitudes

However the normativity of requirements of structural rationality is explained, these requirements deliver conclusions about agents' reasons only by grounding these conclusions in attitudes that those agents already have. As Korsgaard put it, claims about the reasons an agent has must be grounded in things that are "already true of" the agent.[3] This leads us to the question of what attitudes can play this role of anchoring conclusions about the reasons an agent has.

One obvious answer would be that these must be attitudes that are themselves supported by good reasons. If I have good reasons for aiming to be in Chicago by morning, then the fact that I need to leave for the airport now in order to do that is in fact a reason to leave for the airport now. But this alternative is not available for the strategy for explaining particular truths about reasons that I am now considering, since it would involve appealing to such truths at the outset. So some alternative is needed.

Some Kantians appeal here to a version of the first idea of rationality mentioned previously, the idea of a rational agent, and hold that there are certain attitudes that every rational agent is required to have (Korsgaard 1996: Lecture 3). I do not find these arguments persuasive. But even if successful, they would account for only a subset of the reasons people seem to have, basically those expressed in moral requirements, broadly understood. Such an account would therefore need to be supplemented in some way. Korsgaard does this by allowing for reasons individuals have in virtue of the identities that they have adopted, including such things as roles and relationships and professions (Korsgaard 1996; Korsgaard 2009: 20).

[2] See (Raz 2005; Raz 2009; Kolodny 2005; Kolodny 2011) and further references therein.

[3] The range of views about reasons that appeal to such a link is very broad. It includes Humeans such as Bernard Williams, who held that a person has a reason to do something only if this conclusion follows by a "sound deliberative route" from that person's subjective motivational set (1981), and also Kantians such as Korsgaard, who believe that reasons for action must derive from that person's own will (Korsgaard 1996: 19). It is, I believe, the thread that links Michael Smith's Humean view (1994) with his more recent "constructivism" (2013).

Adopting an identity in this sense is something like adopting a large-scale intention or plan of life. As I have pointed out, such choices can affect the reasons one has. But this cannot be a *basic* source of reasons since, again, decisions of this kind generate reasons only if they are themselves supported by such reasons.

Another set of views hold that a person has reason to do what will promote the satisfaction of his or her desires or is in accord with his or her evaluative attitudes. Subjectivist views of this kind have been held within philosophy and even more widely outside of it. So the questions that need to be addressed include:

> *(6)* Why is a subjectivist account, according to which a person's reasons for action depend on that person's desires or other evaluative attitudes, appealing? Should such an account be accepted?

It may seem obvious that the reasons a person has depend on his or her desires. This idea is plausible when we are thinking of reasons in the purely psychological, explanatory sense of "the reason why a person did what she did." Desires do provide reasons in this psychological sense, as in "He did it because he wanted to go to Chicago." But this sense of "a reason" is quite different from the normative sense with which I am now concerned, the sense in which one asks, "Did he have any reason to go to Chicago?"

Even when we focus on reasons in this normative sense, however, the idea that the reasons a person has depend on his or her desires remains appealing. After all, it may seem clear that people can have reason to do different things, even in the normative sense, because they have different desires (as emphasized in Schroeder 2007). Perhaps surprisingly, however, the idea that desires provide reasons is not so plausible from an agent's own point of view. To desire something involves seeing it as having some feature that makes it attractive, such as the fact that one would enjoy it. It is this fact, rather than the desire itself, that seems on reflection to be a reason, from the agent's point of view. The desire is simply a state of regarding this fact as a reason. Differences in the normative reasons that different agents have can be explained by differences in what they have reason to see as desirable, such as by differences in what they will enjoy. (It is differences in individuals' reasons in the psychological sense that are explained by different desires.) Considerations of this kind have led me to conclude that desires never provide reasons for action in the way that desire theories maintain.[4] This view is quite controversial, however.

Some theories that base normative reasons on desires simply identify facts about a person's reasons for action with psychological facts about that person's desires and facts about what will lead to the satisfaction of these desires (Schroeder 2007). A reductive view of this kind would provide a systematic account of reasons for action that avoids metaphysical worries about non-normative facts and properties. In my view,

[4] See (Scanlon 1998: 41–49) for a defense of this claim. (Chang 2004) defends the more moderate

it would do this at the cost of eliminating the normativity of facts about reasons.[5] This seems to me a bad bargain, since, as I have argued previously, metaphysical worries about non-normative facts and properties are misplaced.

Normative desire theories, by contrast, start with an avowedly normative general thesis that agents have reason to do what will fulfill their desires or is in accord with their other evaluative attitudes. Although the idea that desires provide reasons for action has natural intuitive appeal, most contemporary defenders of views of this kind base reasons on a wider range of attitudes. Bernard Williams, for example, says that an agent's reasons depend on his or her "subjective motivational set," which includes such things as "dispositions of evaluation, patterns of emotional reaction, personal loyalties, and various projects, as they might be called, embodying commitments of the agent" (1981: 105). Sharon Street also holds that a person's reasons for action depend on the full range of his or her "evaluative attitudes," which she emphasizes includes much more than desires as normally conceived (2013: 42–44). Michael Smith states his view in terms of desire but makes clear that he is using this term in a very broad sense (1987: 54).

These theories are appealing for a number of reasons. For one thing, they offer an overall account of truths about reasons for action, thus supporting the idea that claims about reasons generally have determinate truth values. Another important source of the appeal of desire theories lies in the general strategy we are now discussing: the idea that the normative authority of reasons for action needs to be grounded in something about the agent. Insofar as evaluative attitudes, like desires, are not subject to a person's will, the idea that these attitudes can anchor conclusions about reasons would not lead to implausible "bootstrapping" in which a person can generate reasons simply by making certain choices, such as by adopting a plan or intention.[6]

These views can still lead to implausible conclusions about the reasons agents have, insofar as agent's evaluative attitudes can be silly or even perverse, such as Caligula's desire to torture people or his evaluative view that their suffering would be a good thing. Some implications of this kind can be avoided by claiming that a person's reasons for action are determined not by just any desire or evaluative attitudes that he or she happens to have but rather by those attitudes that would survive reflective criticism of some kind or ones that the person would come to have under some other ideal conditions. Smith, for example, held that reasons depend on what an agent would desire for him- or herself under present conditions if he or she were fully rational (1994: 151ff).

If, however, the attitudes that a person would have under more ideal conditions are different from those that the person currently has, this raises the question of why it

---

view that desires sometimes provide reasons for action, although not all reasons are based on desires.

[5] (Scanlon 2014: 42–50) For defenses of an alternative view according to which there are normative concepts but only naturalistic properties, see (Schroeder 2007: 79ff) and Gibbard 2003: 29–34).

[6] The "commitments" that Williams mentions might be an exception if these can be adopted at will and do not need to be grounded in other reasons.

is those desires that determine that agent's reasons. It might be suspected that the answer must be that these are conditions under which a person is more likely to desire things that he or she actually has reason to want.

But the desire theorist could say instead that the process of idealization is a matter of working out what the agent really favors rather than one of discovering what is really desirable. Williams, for example, says that an agent has a reason to do something if this conclusion could be reached from the agent's current subjective motivational set by means of a "sound deliberative route" (1981: 104). Williams emphasizes that this process of deliberation can involve modifying and even eliminating some current desires. But the agent's current evaluative attitudes are modified only in the light of other such attitudes that the agent currently holds. Conclusions about reasons that are arrived at in this way thus retain their basis in what is "already true of the agent." Street, who holds that an agent has those reasons that would be supported by an ideally coherent rendering of his or her current evaluative attitudes, emphasizes the importance of this link with the agent's contingent normative starting points (2013: 41). Retaining this link comes, however, at the price of limiting the degree to which these views can avoid seemingly implausible conclusions. As Street seems to acknowledge, such a view may lead to the conclusion that Caligula had good normative reason to torture people, since even an "ideally coherent Caligula" would see himself as having such reasons (2009: 294).

I observed earlier that a theory according to which desires provide reasons to do what will promote their fulfillment does not fit well with the outlook of an individual agent. Desiring something involves seeing something about it as desirable, such as that it would be pleasant, and it is this feature of the thing desired, rather than the desire itself, that the agent sees as providing a reason for action. This is even more evident when we shift from desires to "evaluative attitudes," since these attitudes even more clearly involve judging things to be valuable, or to be promoted, because of certain properties that they have. It is these properties, rather than the agent's own attitudes, that the agent who has these attitudes sees as providing reasons. So there is a gap between the subjectivist theory that identifies an agent's attitudes as the source of his or her reasons and the outlook of these agents themselves.

This gap is closed, in Street's view, by the fact that any ideally coherent agent will come to believe that a realist interpretation of the reasons she has is untenable, for the epistemological reasons I mentioned previously. For any agent, she says, her subjectivist view will therefore be the only account of reasons that withstands critical scrutiny (2016: 331). But even if Street's epistemological argument provided conclusive reason to reject normative realism, it is not clear why it would provide positive support for the normative thesis on which her subjectivist view rests rather than leaving us with a form of nihilism (Berker 2017).

4 The relation between normative truths and non-normative truths

I will conclude with some questions about what is commonly called the divide between facts and values – in more general terms, the distinction between normative and non-normative truths. The relation between these two domains can seem puzzling. On the one hand, it is widely believed (except by reductive naturalists) that normative statements are not logically or conceptually tied to non-normative statements. But it is obvious that what is the case normatively speaking (what reasons people have) depends on what the non-normative facts are, and that the normative facts vary when non-normative facts are different. This is often put by saying that normative facts "supervene on" the non-normative facts, which means that it is a necessary truth (in some sense of "necessary") that normative facts cannot vary as long as the non-normative facts remain the same.[1] It thus seems puzzling why these two realms of facts should be linked in this way if there is no logical or conceptual tie between them. So the questions are:

> *(7)* How should the relation between normative truths and non-normative truths be understood? How is the dependence of the former on the latter to be explained?

Here I will just briefly state my own answer, which is that what appears to be a puzzle here results from misunderstanding the character of normative claims (Scanlon 2014: 40–41).

First, although normative claims, which I will understand to be claims about reasons, are not reducible to non-normative ones, it is not the case that no normative claim is entailed by non-normative ones. The sharpness of this knife is a reason for me not to press my hand against its edge only if the knife is in fact sharp. So from the (non-normative) fact that the knife is not sharp (that is to say, not sharp enough to penetrate my flesh easily), it follows that that the normative claim that the sharpness of the knife is a reason for me not to press my hand against its edge is in fact false. The normative claims that are not logically tied to non-normative ones are just what I call "pure normative claims," such as the claim that if the knife were sharp enough to easily penetrate my flesh (and as long as certain other non-normative conditions

[1] See (Dreier 1992) for discussion.

also held), the sharpness of the knife would be a reason for me not to press my hand against its edge. Pure normative claims of this kind are a small subset of all normative claims, most of which are "mixed" claims like the one I began with.

Second, not all normative truths vary as the non-normative facts vary. Only the truth values of *mixed* normative claims vary in this way. The truth of pure normative truths does not, because they do not depend on non-normative facts.

Third, what pure normative truths do is to specify the way in which the truth of mixed normative claims depends on non-normative facts. So the relation between the normative and the non-normative is a *normative* matter. This relation is therefore very different from the relation between mental facts and physical facts, which is a common example of a relation of supervenience. The content of the most basic normative claims (the pure normative claims) is precisely to assign normative significance to possible non-normative facts (for example, to say whether, should they obtain, certain facts would constitute reasons.) The most basic claims about mental phenomena do not have this role.

Fourth, whether normative facts supervene on the non-normative facts in the usual technical sense depends on whether pure normative truths are metaphysically necessary or necessary in some other sense. This strikes me as a metaphysical question on which it does not seem to me necessary to take a stand in order to explain away what initially seemed puzzling about the relation between the normative and the non-normative.

5 A very brief reiteration

I have identified seven questions, or sets of questions, that students interested in practical reason should address and surveyed possible answers to them, indicating which ones I favor. These questions are mostly familiar ones. My main aim in identifying them has been not just to say that they need to be answered but to say that they should not be taken at face value: to call attention to some further questions about the ways in which these questions are commonly formulated and to some presuppositions that lie behind them.

References

Bedke, Matthew 2009, "Intuitive Non-Naturalism Meets Cosmic Coincidence," *Pacific Philosophical Quarterly* 90, 188–209.

Benacerraf, Paul 1973, "Mathematical Truth," *The Journal of Philosophy* 70, 661–679.

Berker, Selim 2017, "Does Evolutionary Psychology Show That Normativity Is Mind-Dependent?" In Justin D'Arms and Daniel Jacobsen (eds.) *Moral Psychology and Human Agency: Essays on the New Science of Ethics*, Oxford: Oxford University Press.

Broome, John 2007, "Does Rationality Consist in Responding Correctly to Reasons?" *Journal of Moral Philosophy* 4, 349–74.

——— 2013, *Rationality Through Reasoning*, Oxford: Wiley-Blackwell.

Chang, Ruth 2004, "Can Desires Provide Reasons for Action?" In R. Jay Wallace, Philip Pettit, Samuel Scheffler, and Michael Smith (eds.) *Reason and Value: Themes from the Moral Philosophy of Joseph Raz*, New York: Oxford University Press, 56–90.

Clarke-Doane, Justin 2017, "What Is the Benacerraf Problem?" In Fabrice Pataut (ed.) *Perspectives on the Philosophy of Paul Benacerraf: Truth, Objects, Infinity*, Dordrecht: Springer.

Davidson, Donald 1980, "Actions, Reasons, and Causes," in Donald Davidson (ed.) *Essays on Actions and Events*, Oxford: Clarendon Press, 3–19.

Dreier, Jamie 1992, "The Supervenience Argument Against Moral Realism," *Southern Journal of Philosophy* 30, 13–38.

Gibbard, Allan 2003, *Thinking How to Live*, Cambridge, MA: Harvrd University Press.

Hare, R. M. 1952, *The Language of Morals*, Oxford: Clarendon Press.

Joyce, Richard 2001, *The Myth of Morality*, Cambridge: Cambridge University Press.

Kolodny, Niko, 2005, "Why Be Rational?" *Mind* 114, 509–563.

——— 2011, "Aims as Reasons" in Samuel Freeman, Rahul Kumar, and R. Jay Wallace (eds.) *Reasons and Recognition: Essays on the Philosophy of T. M. Scanlon*, New York: Oxford University Press, 43–78.

Korsgaard, Christine 1996, *The Sources of Normativity*, Cambridge: Cambridge University Press.

_____ *2009,* Self-Constitution: Agency, Identity, and Integrity*, Oxford: Oxford University Press.*

Mackie, John 1977, *Ethics: Inventing Right and Wrong*, Harmondsworth: Penguin Books.

McDowell, John 1985, "Values and Secondary Qualities," in Ted Honderich (ed.) *Moralty and Objectivity*, London: Routledge & Kegan Paul, 110–129.
Parfit, Derek 2011, *On What Matters*, Vol. I and II, Oxford: Oxford University Press.
Raz, Joseph 2005, "The Myth of Instrumental Rationality," *Journal of Ethics and Social Philosophy* 1, 1–28.
——— 2009, "Reasons: Practical and Adaptive," in David Sobel and Steven Wall (eds.) *Reasons for Action*, Cambridge: Cambridge University Press.
Scanlon, T. M. 1998. *What We Owe to Each Other*, Cambridge, MA: Harvard University Press.
——— 2007, "Structural Irrationality," in Geoffrey Brennan, Robert Goodin, Frank Jackson, and Michael Smith (eds.) *Common Minds: Essays in Honor of Philip Pettit*, Oxford: Oxford University Press, 84–103.
——— 2014, *Being Realistic About Reasons*, Oxford: Oxford University Press.
——— 2017, "Normative Realism and Ontology: Reply to Clarke-Doane, Rosen, and Enoch and McPherson," *Canadian Journal of Philosophy* 47, 877–897.
Schroeder, Mark 2007, *Slaves of the Passions*, Oxford: Oxford University Press.
Skorupski, John 2010, *The Domain of Reasons*, Oxford: Oxford University Press.
Smith, Michael 1987, "The Humean Theory of Motivation," *Mind* 96, 36–61.
——— 1994, *The Moral Problem*, Oxford: Blackwell Publishers.
——— 2013, "A Constitutivist Theory of Reasons: Its Promise and Parts," *Law, Ethics and Philosophy* 1, 1–30.
Street, Sharon 2006, "A Darwinian Dilemma for Realist Theories of Value," *Philosophical Studies* 127, 109–166.
——— 2008, "Constructivism About Reasons," in Russ Shafer Landau (ed.) *Oxford Studies in Metaethics*, Vol. 3, New York: Oxford University Press, 207–245.
——— 2009, "In Defense of Future Tuesday Indifference: Ideally Coherent Eccentrics and the Contingency of What Matters," *Philosophical Issues* 19, 273–298.
——— 2013, "Coming to Terms with Contingency: Humean Constitutivism About Practical Reason," in James Lenman and Yonathan Shemmer (eds.) *Constructivism in Practical Philosophy*, Oxford: Oxford University Press.
——— 2016, "Objectivity and Truth: You'd Better Rethink It" in Russ Shafer Landau (ed.) *Oxford Studies in Metaethics*, Vol. 11, New York: Oxford University Press, 293–334.
Tait, W. W. 1986, "Truth and Proof: The Platonism of Mathematics," *Synthese* 69, 341–370.
_ _ _ _ _ _ _ *2005,* The Provenance of Pure Reason: Essays in the Philosophy of Mathematics and Its History*, Oxford: Oxford University Press.*
Williams, Bernard 1981, "Internal and External Reasons" in Bernard Williams (ed.) *Moral Luck*, Cambridge: Cambridge University Press, 101–113.

2. Practical reason

Rationality or normativity but not both
John Broome

1 Introduction

The term ‘practical reason’ has two quite different meanings. They arise from the ambiguity of the noun ‘reason.’ In one of its meanings, practical reason can be identified with rationality and in the other with normativity. But the two meanings are often fused in philosophy, and indeed the distinction between rationality and normativity is often obscured. There are philosophical theories that identify the two. This chapter aims to hold apart the two meanings of ‘practical reason.’

It starts by analyzing the rational meaning and the normative meaning in sections 2 and 3, respectively. Section 4 gives a simple reason for thinking that rationality and normativity must be distinct: one is a mental property, and the other is not. Your degree of rationality depends only the properties of your mind, whereas often you ought normatively to act on the world outside your mind. Section 5 examines and rejects two possible responses. One denies that rationality is a mental property; the other asserts that normativity is a mental property. Section 6 describes a different argument in a Kantian spirit. A special, reified concept of Rationality might possibly be united with normativity, but it is far from our ordinary concept of rationality. Section 7 summarizes the chapter’s argument and draws the conclusion that the two meanings of ‘practical reason’ cannot be successfully unified.

This book therefore covers two distinct topics.

2 Rationality

In one of its senses, 'reason' is the name of a property that is possessed to a greater or lesser degree by people. This property can be ascribed to someone by means of the adjective 'rational,' which corresponds to the noun 'reason' in this sense. To say a person is rational is just to say that she has reason in this sense.

The noun 'rationality' is a name for this property, formed from 'rational' by attaching the noun-forming suffix '-ity.' In the course of history, the noun 'reason' gave rise to the adjective 'rational,' which in turn gave rise to the new noun 'rationality.' The *Oxford English Dictionary* (2006) records 'reason' in this sense from 1225, 'rational' from 1398 and 'rationality' (in the non-mathematical sense) from 1628. 'Rationality' is therefore strictly speaking a redundant word, since we already have 'reason' as a name for the property that is ascribed by 'rational.' But 'rationality' is useful all the same because it is less ambiguous than 'reason.' It is synonymous with 'reason' in just one of its senses – the one I am describing in this section. I shall often use 'rationality' in place of 'reason' in this sense. I shall also use the expression 'reason in the rational sense.' This terminology helps to limit the confusion that can be caused by the ambiguity of 'reason.'

When 'reason' has the rational sense, 'practical reason' means the same as 'practical rationality.' It refers to the part of rationality that is concerned with action. What does that include? What is the subject matter of the philosophy of practical reason in this sense?

First, the subject matter is mental. A conceptual feature of rationality is that a person's rationality is an intrinsic property of her mind, and it supervenes on other intrinsic properties of her mind. If your mind has exactly the same intrinsic properties (apart from rationality) in one possible state as it has in another, then you are exactly as rational in the one as you are in the other. Rationality supervenes on the mind, as I shall put it, taking this expression from Ralph Wedgwood (2002). And I shall use 'mental property' to refer to an intrinsic property of the mind.

For example, when a person intends to drink a glass of liquid, she is equally rational in the case when the liquid is petrol as she is in the case when it is gin, so long as the difference is not registered in any mental property of hers.[1] For another example, if you fail to take a means to an end that you intend, this is not necessarily a failure of rationality if it is caused by some non-mental obstruction. It could be that, if you had had all the mental properties you do have, but the obstruction had been removed, you

[1] This famous example comes from Bernard Williams (1981).

would have taken the means to your end. There would then have been no failure in your rationality, so supervenience implies there is no failure in the actual case either.

We sometimes ascribe the property of rationality to other things besides a person. For example, a belief may be said to be rational. So may an act, a government's policy, or even a city plan. The primary bearers of rationality are nevertheless people. Ascriptions of rationality to other things are derivative in one way or another. For example, to say a person's action was rational may mean that, had she not done it, she would have been no more rational than she is, having done it. Or it may mean that her action constitutes evidence of her rationality. To say a city plan is rational may mean that it could have been designed by a rational person. These derivative sorts of rationality are not necessarily mental properties. But the rationality of a person is a mental property of hers.

What about mental externalism? I believe the Taj Mahal is made of marble. Suppose that, elsewhere in the universe, there is a Twin Earth that has all the same intrinsic physical properties as Earth. On Twin Earth lives a person called 'John Broome.' His intrinsic physical properties, including the intrinsic properties of his brain, are the same as mine. He has a belief that he would express with the words 'The Taj Mahal is made of marble.' But he has no beliefs about the Taj Mahal, because the Taj Mahal is on Earth and he has no beliefs about Earth. His belief is about the Twin Taj Mahal. My belief is about the Taj Mahal, so his belief is not the same as mine, since it has a different content and beliefs are individuated by their contents. At least, that is the view of mental externalists. I take each of a person's beliefs to be a mental property of hers. So Twin John's mind does not have all the same mental properties as mine has. If externalism is true, our mental properties do not supervene on our intrinsic physical properties, therefore.

For an analogy, think of the magnetic field of a particular magnet. The field belongs to the magnet, but it does not supervene on the magnet's intrinsic physical properties, since other ferrous objects in the neighborhood influence the field. The field's direction and strength at a particular point do not supervene on intrinsic physical properties of the magnet. However, they are intrinsic properties of the magnet's field; indeed, the field simply consists of the set of directions and strengths at all points. The analogy is this: a person corresponds to a magnet; the person's mind corresponds to the magnet's field; a belief corresponds to the direction and strength of the field at a point. A person's belief supervenes on the intrinsic properties of her mind, but it does not supervene on her intrinsic physical properties

If externalism is correct, the principle that rationality supervenes on the mind does not imply that Twin John and I are exactly as rational as each other, since our minds are different. Nevertheless, we surely are exactly as rational as each other, so presumably there is some stronger principle that does have this implication. Presumably it would be a principle that rationality supervenes on internal properties of the mind, defined in some way or other. But I do not know any such principle, and I do not affirm that one exists. At any rate, externalism is no threat to the principle I do affirm,

that a person's rationality supervenes on her mental properties apart from rationality itself. It simply suggests there may also be a stronger principle.

Some acts are mental, such as acts of doing mental arithmetic, but many are not. If rationality is a mental property, how can it be concerned with non-mental acts? How can it be practical? The answer is that some mental properties are intimately related to action. Intentions are the prime example. A necessary condition for doing an act is that you have a particular intention. Also, an intention generally causes an act, and it is the object or purpose of an intention to do so. Practical rationality is primarily the part of rationality that is concerned with intentions.

The property of rationality – of reason in the rational sense – is often described more specifically as an ability or a capacity or a faculty. But that description is too narrow. Suppose you intend some end but do not intend an act that you believe to be a necessary means to this end. This is a lapse in your rationality, whatever abilities you may have. If you have the ability to acquire the intention of doing this act, well and good, but your rationality is still lacking unless you exercise this ability. Moreover, your rationality might improve even without your exercising an ability. You might come to intend an act that you believe is a necessary means to an end you intend not by doing anything at all but as a result of some subpersonal process within you. You are then more rational than you were before, even though you have not exercised your rational ability. One part of rationality is concerned with a person's mental states and the relations between them rather than with her abilities.

Nevertheless, an important part of being rational is having a rational ability, and another part is the process of exercising it. A rational ability is commonly exercised through the activity of reasoning, so reasoning falls under rationality. The philosophy of practical reason includes practical reasoning within its subject matter; several chapters in this book are concerned with it. Furthermore, the correctness of reasoning is also a part of rationality. Any creature that reasons is rational to an extent, but a creature that reasons correctly is more rational than one that reasons incorrectly.

We sometimes reify the property of rationality, treating it as a thing rather than a property. We do the same for morality. Morality is the property a person possesses when she is moral, but sometimes we treat it as something that stands outside a person and has or at least pretends to some sort of authority over her. Similarly we sometimes treat rationality – reason in the rational sense – as something that stands outside a person and has some sort of authority over her. The philosophy of rationality encompasses reified rationality.

Reification is most apparent in the expression 'rationality requires.' 'Rationality requires you to intend means to your end' may mean simply that intending means to your end is a necessary condition for your having the property of rationality. With this meaning, rationality is not reified. The sentence follows the model of 'Survival requires you to drink water,' which means that your drinking water is a necessary condition for your having the property of survival. However, 'Rationality requires you to intend means to your end' may alternatively be understood on the model of 'The

law requires you to pay taxes.' This latter sentence means that the law prescribes that you pay taxes. Similarly, 'Rationality requires you to intend means to your end,' when rationality is reified, means that rationality prescribes that you intend means to your end.

I claim that the expression 'rationality requires' is most naturally understood in the reified sense.[2] I think we would not naturally say that rationality requires you to be alive. But being alive is a necessary condition for possessing the property of rationality, so unreified rationality does indeed require you to be alive. On the other hand, rationality does not prescribe that you are alive, so reified rationality does not require you to be alive. We naturally take 'rationality requires' this second way. What rationality requires in this more natural sense is a subset of the necessary conditions for being rational.

All of these aspects of rationality, when applied to the practical sphere, provide the subject matter of the philosophy of practical reason. What exactly does rationality require of our intentions? What is practical reasoning, exactly, and what distinguishes correct from incorrect practical reasoning? These are questions for the philosophy of practical reason. An important question is whether there is a distinctly practical branch of rationality at all. For example, it has been argued that rational requirements on intentions are only derivative from rational requirements on beliefs (e. g. Ross 2009, Setiya 2007, Wallace 2001). David Hume (1988: 458) famously denied that there is any such thing as practical reasoning: 'Reason is the discovery of truth or falsehood,' he said.

[2] My evidence is contained in (Broome 2013: 119–126). Here I give just one example.

3 Normativity

A different sense of 'reason' gives us the second meaning of 'practical reason.' 'Reason' in this different sense is a mass noun. It is derived from the count noun 'a reason,' whose plural is 'reasons.' The *Oxford English Dictionary* (2006) gives an example of the count noun from 1225, but no clear examples of the mass noun earlier than Shakespeare.

The count noun itself has various senses, which philosophers have identified as a motivating sense, an explanatory sense, a normative sense and so on. I use it in the normative sense only. In this chapter, 'a reason' always refers to a normative reason.

Count nouns are often associated with corresponding mass nouns. For example, 'a light' is associated with 'light' and 'a pleasure' with 'pleasure.' Often the things referred to by the count noun explain or give rise to whatever is referred to by the mass noun (Fogal 2016). We often use simply 'give' for this relation. A light gives light and a pleasure gives pleasure. Similarly, the count noun 'a reason' is associated with the mass noun 'reason.' A reason gives reason. For example, the fact that a restaurant is noisy, which is a reason to avoid it, gives reason to avoid this restaurant.

Grammatically, a mass noun refers to stuff of some sort. Whereas 'light' refers to a sort of physical stuff and 'pleasure' refers to a sort of mental stuff, it is obscure what the mass noun 'reason' refers to. Grammatically, it should refer to normative stuff of some sort. However, it seems metaphysically improbable that such normative stuff really exists.

The metaphysics of reason makes little difference to this chapter, but it deserves a mention. I think the mass noun does not refer to anything. I think that saying 'There is reason to avoid that restaurant' is merely the way we have in English of ascribing a particular normative property to the act of avoiding that restaurant. English provides no adjective for the property that something has when there is reason for it. You might think 'reasonable' is such an adjective, but it is not. Even when something has the property of being something there is reason for, it may not be reasonable. There may be reason to avoid a restaurant, but stronger reason not to avoid it. If so, avoiding the restaurant would not be reasonable. In the absence of an adjective, we can ascribe the property only by using the expression 'there is reason.' This expression apparently refers to stuff but does not really do so; it ascribes a property.

In this chapter, it does not matter whether the mass noun 'reason' refers to something. What matters is that there is a sense of 'reason' in which it is a mass noun and has a normative meaning. To contrast it with the rational sense of 'reason,' I call this the 'normative sense.' Reason in this sense is closely associated with reasons. Reason is

given by reasons, as light is given by lights. Whenever there is reason, something gives rise to it, and that something is a reason. Whenever there is a reason, it gives rise to reason. The sentences 'There is a reason for A to F' and 'There is reason for A to F' are therefore true or false together. Consequently, many philosophers slip between the mass noun 'reason' and the count noun 'a reason,' apparently without noticing (e.g. Nagel 1970; Skorupski 2010; Smith 2004).

Those two sentences actually have significantly different meanings. 'There is a reason for A to F' is a quantified statement, which asserts the existence of something in the world that is a reason for A to F. 'There is reason for A to F' means only that A's Fing has the particular normative property of being something there is reason for; it does not assert the existence of anything. At least, that is my view. Still, so long as metaphysics in not in question, slipping between reason and reasons can be harmless. A philosopher who writes of practical reason in the normative sense might equally well write of practical reasons.

Indeed, when 'reason' has the normative sense, 'practical reason' may be used to refer to all of practical normativity. The philosophy of practical reason in this sense is concerned with everything that reasons require of us in practical matters. It is concerned with what our reasons require us to do or to intend.

Reasons come from various sources: morality, prudence and the law are among them. Rationality may also be among them: it may be that if rationality requires something of you, this fact constitutes a reason for you to achieve what it requires.[1] So morality, prudence, the law and perhaps rationality all contribute to normativity; they help to determine what our reasons require of us.

What your reasons require of you is, put differently, what you ought. Here, I use 'ought' in the 'all-things-considered' or 'final' sense, as I shall throughout this chapter. Philosophers sometimes use 'ought' for what morality in particular requires of you; I do not use it that way.

To make a closer parallel with 'rationality,' I shall treat 'normativity' as the name of a property than can be possessed by a person. 'Normative' is its corresponding adjective. Normativity is the property you have when you do what you ought to do, believe what you ought to believe, hope for what you ought to hope for and so on. The property of normativity is the property of Fing when you ought to F. This property comes in degrees: the more you perform as you ought, the more normative you are.

This use of 'normativity' and 'normative' is an innovation. But it should not be offensive, because 'normativity' as moral philosophers use the word is an artificial term anyway. It should therefore not be offensive to treat it artificially as a property of people.

I shall further cement the parallel between rationality and normativity by using the term 'normative requirement' for an ought. When you ought to do something, I shall sometimes say that normativity requires you to do it.

[1] It is a difficult question whether or not this is so. See, for example, Broome (2013: chapter 11),

Southwood (2010), Lord (2010).

4 Rationality is not the same as normativity

I have identified two different meanings of 'practical reason': practical rationality on the one hand and practical normativity on the other. They arise from quite different senses of 'reason.' Indeed, 'reason' in these two senses belongs to two different grammatical categories: in one sense, it is the name of a property and in the other a mass noun.[1] The two meanings of 'practical reason' should not be confused.

When they write about practical reason, some philosophers deliberately adopt the normative meaning. But the chapters in this book show that philosophers more commonly adopt the rational meaning and set out to write about practical rationality including practical reasoning. In the *Stanford Encyclopedia of Philosophy*, Jay Wallace (2020) defines practical reason as 'the general human capacity for resolving, through reflection, the question of what one is to do.' This definition is too narrow, since rationality is more than a capacity, but at any rate, it picks out the rational meaning of 'practical reason' rather than the normative one. However, although their aim is to write about rationality, many philosophers end up discussing reasons, which are a feature of normativity rather than rationality. They may have a justification for associating normativity with rationality in this way; the rest of this chapter discusses some potential justifications. But it should not happen through accidental equivocation between the rational meaning of 'practical reason' and the normative one. The normative meaning is slightly obscure, since it comes from the mass noun 'reason,' whose reference is obscure, but it is a bona fide meaning that makes it correct to discuss reasons under the heading 'practical reason.' The mistake is to confuse the two meanings.

Still, there are articulated philosophical views that identify rationality with normativity (e.g. Lord 2017).[2] Although there are two different meanings of 'practical reason,' involving two different grammatical categories, some philosophers think the difference does not represent any real distinction. They think the two grammatical routes I mapped out through different senses of 'reason' arrive eventually at the same destination.

Many philosophers suppose that your rationality is a matter of achieving what reasons require of you. More precisely, they think that rationality consists in *F*ing

[1] My thanks to Rob Bassett for pointing this out to me.

[2] Lord's (2018) book *The Importance of Being Rational* has been published since this paper was written. Lord's views have developed, and my remarks here may no longer be entirely apposite.

when your reasons require you to *F*. But this is *F*ing when you ought to *F*, which is just what normativity is. Their supposition is that the property of rationality is the same as the property of normativity as I defined it. They treat rationality and normativity as identical. Let us call their view 'the identity theory.'

We might be less interested in whether the properties of rationality and normativity are the same than in whether what rationality requires of you is the same as what normativity requires of you. Let us give the name 'the requirement-identity theory' to the theory that rationality requires a person to do something if and only if normativity requires her to do it, which is to say if and only if she ought to do it. The requirement-identity theory implies the identity theory. To see why, suppose the requirement-identity theory is true, and suppose you are rational. Because you are rational, you satisfy all the rational requirements you are under. The requirement-identity theory tells us that your rational requirements are the same as your normative requirements, so you also satisfy all the normative requirements you are under. You are therefore normative. So, necessarily, if you are rational, you are normative. By a parallel argument, necessarily, if you are normative, you are rational. Properties that necessarily have the same extensions are identical properties, or so I shall assume. So the requirement-identity theory implies the identity theory. I intend to refute both theories. For this reason, I shall concentrate on the identity theory, since a refutation of this theory is also a refutation of the other.

There are definitely strong connections between rationality and normativity. I said in section 3 that rationality may well be a source of normativity; if rationality requires something of you, that may well be a reason for you to achieve what it requires. Furthermore, rationality requires you to intend to do whatever you believe you ought to do. I call this requirement 'enkrasia' (Broome 2013: section 9.5) It requires you to respond to your normative beliefs in a particular way. This is only one of many practical requirements of rationality; for instance, another is the requirement to intend whatever you believe is a means implied by an end you intend. Still, it does constitute an important connection between rationality and normativity.

But rationality is not identical to normativity. There is a prima facie objection to the identity theory; I call it 'the quick objection' (Broome 2013: section 5.2). It is simply that rationality supervenes on the mind and normativity does not. If your mental properties (apart from rationality) are the same in one possible state as they are in another, you are equally rational in the two states, but you may not be equally normative. Here are two examples.

One is the drinks example again. You intend to drink a glass of liquid. In one case the liquid is petrol and in the other gin, but assume the difference does not register in any mental property of yours. To complete the story, suppose the glass is unexpectedly taken away from you before you carry out your intention to drink, so you never find out what is in it. The difference between the two cases makes no difference to your mind at any time. You are equally rational in either case because rationality supervenes on the mind. However, if the circumstances are normal, in the first case you ought not to

intend to drink the liquid because drinking it is harmful, whereas in the second case this is not so. Since you intend to drink the liquid, in the first case, you are not as you ought to be in this respect, whereas in the second case, that is not so. You may be normative in the second case but not in the first.

Here is the second example. You ought to insure your house against fire. Moreover, you believe this is so, and you set about insuring your house. You complete an application form and pay a premium to an insurance company in the usual way, without having studied all the fine print carefully. Now take two different cases. In the first, everything proceeds as expected, and your house is insured. In the second case, the small print contains a clause that says your house is insured only if its roof is constructed of slate, tiles or metal. Actually your house's roof is constructed of cedar shingles, so the house is not insured. Suppose this fact never comes to your attention because there is no fire. Then your mental properties are exactly the same in both cases; in particular, in both cases you believe you insure your house. You are therefore equally rational in both cases. Yet in one case you insure your house as you ought, and in the other you do not. You may be normative in one, but you are not normative in the other.

5 Responses

The quick objection to the identity theory can be opposed in only two ways. One is by asserting that normativity supervenes on the mind. The other is by denying that rationality supervenes on the mind. Neither of these responses is sufficient to prove the identity theory, but refuting both is sufficient to refute the identity theory. In this section, I aim to refute both.

Does what you ought supervene on the mind?

The first response to the quick objection is to assert that your normative situation cannot be different in two cases if your mental properties are the same. The argument for this claim breaks into two parts. The first is to argue that what you ought – what normativity requires of you – supervenes on your mind.

One way of making this argument is to claim that all reasons are states of mind. For example, they might be pairs consisting of a desire and a belief. In the drinking example, we may assume that you desire not to drink petrol. If you believe the liquid is petrol, you then have a reason not to intend to drink it, and we may assume you ought not to intend to drink it. This is so whether or not the liquid is actually petrol. Conversely, if you believe the liquid is gin, you have no reason not to intend to drink it. Again, this is so whether or not the liquid is actually petrol.

Just because reasons are states of mind, it does not directly follow that what you ought supervenes on your mind. What you ought does not depend only on what your reasons are; it also depends on how your reasons combine and weigh against each other. But, once we assume reasons are states of mind, we may naturally add the further assumption that what you ought also supervenes on the mind.

We emerge with a fully subjectivist view about what normativity requires. Many philosophers are unwilling to accept this much subjectivism, and there is also a less subjectivist argument for the view that what you ought supervenes on your mind. This argument can be developed from Benjamin Kiesewetter's (2018) account of reasons. Kiesewetter accepts that reasons are often objective and outside the mind; for example, they may be facts about the world. But he thinks that a necessary condition for something to be a reason for you is that it impinges on your mind in some way.[1] More precisely, it is part of your evidence, which Kiesewetter takes to entail that either

[1] More accurately, Kiesewetter takes this to be true of reasons to do something in the present but not necessarily of reasons to do something in the future.

you believe it, or it is evidenced by some more inchoate state of your mind such as a phenomenal experience. He assumes that what you ought is determined by your reasons, which are necessarily part of your evidence.

This is not enough to ensure that what you ought supervenes on your mind. In the drinks example, suppose the liquid is petrol and this fact is part of your evidence. You believe the liquid is petrol, and perhaps you experience associated phenomena such as a particular smell. Then according to Kiesewetter, you have a reason not to intend to drink the liquid. The reason is that the liquid is petrol. Let us assume that consequently you ought not to intend to drink the liquid. Now compare a different case in which the liquid is gin, but your mental properties are exactly as before. You believe the liquid is petrol, and you have the same associated mental phenomena. The phenomena are illusory in this case, of course. In this case, you have no reason not to intend to drink the liquid, since it is not petrol. It is not the case that you ought not to intend to drink it. If the liquid is petrol, you have a reason, and if it is gin, you do not, yet your mind has exactly the same properties in either case. So we do not yet have an argument that what you ought supervenes on your mind.

But less us add some assumptions. Let us assume, first, that what you ought is determined by your evidence as a whole. This means it can be affected by evidence that you have a reason as well as evidence that itself constitutes a reason. Second, let us assume that something is part of your evidence only if you know it. This is Timothy Williamson's (2000: chapter 9) view; Kiesewetter mentions it but does not commit himself to it. Third, let us assume that knowledge is a mental state. This too is Williamson's view (2000: chapter 1). Then, if your mind has all the same properties in one case as in another, the third assumptions ensures that what you know in one case is the same as what you know in the other. The second assumption then ensures that your evidence is the same in either case. Finally, the first assumption ensures that what you ought is the same in either case. What you ought supervenes on your mind, therefore. In effect, the assumption that knowledge is a mental state bridges the gap between your mind and the outside world.

These are strong assumptions, and many philosophers would be unwilling to accept them. But for the sake of argument, let us assume that what you ought supervenes on the mind. It does not follow that normativity supervenes on the mind, as I shall shortly explain.

However, you might think that the requirement-identity theory – the theory that what you ought is the same as what rationality requires of you – does follow. What rationality requires of you supervenes on the mind, so if what you ought also supervenes on the mind, does it not follow that what you ought and what rationality requires of you are the same? It does not. In the insurance example, it may be that you ought to insure your house, whereas rationality requires you to intend to insure your house. This is consistent with assuming that what you ought and what rationality requires of you both supervene on your mind. Suppose, for instance, that you want to preserve

your wealth and you believe that insuring your house is a necessary means to that end, and that this is why you ought to insure your house.

Does normativity supervene on the mind? No

We are assuming for the sake of argument that what you ought supervenes on the mind. Normativity is the property of *F*ing whenever you ought to *F*. To establish that normativity supervenes on the mind, we also have to show that, in cases where you ought to *F*, your actually *F*ing supervenes on your mind. This is the second part of the argument that normativity supervenes on the mind.

The insurance example reveals the problem with it. In that example, I assumed you ought to insure your house. But whether or not you insure your house does not supervene on your mind. It depends on the small print in the insurance contract, which you have not studied. So your normativity does not supervene on your mind. In one case, you do as you ought, and in the other, you do not, even though your mind is the same in both cases.

To argue that normativity supervenes on the mind, we need to argue that any act you ought to do must be one whose performance supervenes on the mind. Some acts satisfy this constraint. Mental acts such as doing mental arithmetic satisfy it. So does any act of bringing it about that your mind has a particular property. This need not be a mental act, because you might bring about the property by non-mental means. For instance, by opening a window, you might bring it about that you believe the window is open.

To make the argument in the insurance example would involve denying that you ought to insure your house. Instead, what you ought would have to be something that supervenes on your mind. For instance, it might be that you ought to bring yourself to believe you have insured your house. You can achieve this by the non-mental means of filling in the right form, paying the premium and so on. Alternatively, it might be that you ought to intend to insure your house or to have some other mental property.

Acting on the world outside your mind does not supervene on your mind. For example, raising your arm does not supervene on your mind. You might fail to raise your arm even while your mind has exactly the properties it would have if you raised it. Your nerves might fail to activate your muscles, and you might be looking the other way.

The argument has to be that no act on the world outside your mind can be something you ought to do. That is utterly implausible. The fact is that you ought to insure your house and not merely intend to insure it or believe you have insured it. You ought to make sure your car's brakes are in good condition, you ought to be kind to strangers,

look both ways before you cross the road and so on. It is quite implausible that all these ordinary normative claims are false.[2]

I conclude that normativity does not supervene on the mind.

Does rationality supervene on the mind? – Yes

Errol Lord (2017) favors the identity theory. He adopts the first response to the quick objection; he denies that rationality supervenes on the mind. He thinks (Lord 2017) that rationality consists in responding correctly to the reasons you possess, he thinks that you possess a reason just when you are in a position to know it (Lord 2010), and being in a position to know a reason, as he understands it, does not supervene on the mind.

Lord uses the drinks example to illustrate possessing a reason as he understands it. You intend to drink a glass of liquid, which is either petrol or gin. Lord adds the assumption that, lying on the counter right in front of you, is an authoritative card that specifies which the liquid is. If it is petrol, the fact that it is petrol is a reason not to intend to drink it. The fact that the reason is described on a card right in front of you – and Lord adds some further supporting details – implies that this reason is possessed by you.

However, Lord supposes you do not read the card so that the reason does not impinge on your mind, even though it is possessed by you. You mind has exactly the same properties whether the liquid is petrol or gin, even though in the former case, you possess a reason not to intend to drink the liquid, whereas in the latter case, you do not.

Lord supposes that, if the liquid is petrol, the reason you possess not to intend to drink it outweighs any opposing reasons you might possess, so responding correctly to the reasons you possess entails not intending to drink it. If you intend to drink it, you are therefore irrational according to his theory of rationality. Why should we accept that? Lord evidently thinks he has supplied enough details to convince us that your failure to read the card is a failure of rationality.

Let us accept that that is so. Then you would be equally irrational if the liquid were gin and you intended to drink it. In that case, too, you fail to read the card in front of you. If your failure is irrational in one case, it is irrational in the other. It cannot be otherwise, since your mind is no different in the two cases.

What if the details are such that, in the petrol case, you know something you do not know in the gin case? For example, suppose you know the card contains information that is vital for your health. In the gin case, you do not know this because it is not true. Some philosophers take knowledge to be a mental state (e.g. Williamson 2000:

[2] Kurt Sylvan has pointed out to me that this is precisely what H. A. Prichard (2002: 95–97) thinks. Even with Jonathan Dancy's help, I have not been able to extract a credible argument from Prichard's text.

chapter 1). If they are right, you have different mental properties in the two cases, so the principle that rationality supervenes on the mind does not apply. If they are wrong, we may continue to assume your mind does have the same properties in the two cases, so the principle does apply. We may assume, in particular, that in the gin case, you believe that the card contains information that is vital for your health, just as you do in the petrol case. Given that your failure to read the card is irrational in the petrol case, it is irrational in the gin case too.

Lord's example only makes it clearer that rationality supervenes on the mind. I am speaking of rationality as we ordinarily understand it. There can be other, artificial notions of rationality – 'objective rationality' for instance – that do not supervene on the mind. Kantians often identify normativity with rationality. In section 6, I shall argue that they can do so only by adopting a special, reified notion, which might be called 'Rationality' or 'Reason' with a capital letter. This reified notion need not supervene on the mind. But rationality as we ordinarily understand it does.

I conclude from this section that the quick objection is sound and the identity theory is false.

6 A Kantian response

There remains one way of denying that rationality supervenes on the mind that I have not yet explored. I shall approach it indirectly, by a detour through a puzzle about morality.

A Kantian view is that morality supervenes on the mind. The view is that to be moral, you require only a good mind – specifically, a good will. If, by bad luck, your good will does not achieve good consequences for the world, you are no less moral for that.

Even if morality supervenes on the mind, it does not follow that normativity does so. Morality is only a part of normativity. My examples were about prudence rather than morality, and there is no suggestion that prudence supervenes on the mind. So the conclusions I drew from the examples are not affected. The point I want to make is different. Applied to morality, the Kantian view is plausible, and it raises a puzzle.

It is plausible that your possession of the property of morality does indeed supervene on your mind. If you intend to act well but bad luck intervenes and prevents you from achieving a good result, it is plausible that this does not count against your morality. You cannot be blamed, so you cannot be any less moral. In general, you cannot be more moral in one state than another if your mind is no different. Yet on the other hand, it is implausible that everything morality requires of you – in other words, everything you morally ought – supervenes on your mind. For example, morality requires you not to cause unnecessary suffering, and whether or not you cause unnecessary suffering does not supervene on your mind. Morality is surely aimed at the world, not at improving your own mind. So morality supervenes on the mind, but what morality requires does not. How can these claims be reconciled? That is the puzzle.

A solution is to recognize that the reification of morality has gone a long way. 'Morality' is the name of a property that people possess. Plausibly, this property supervenes on the mind. If we construe 'morality requires' on the model of 'survival requires,' what is morally required of you is whatever is a necessary condition for possessing the property of morality and nothing more. If morality supervenes on the mind, what is morally required in this sense also supervenes on the mind.

But we also reify morality and treat it as an external entity that has some authority over us. Let the capitalized word 'Morality' be the name of the reified entity. Then 'Morality requires' should be construed on the model of 'the law requires' rather than on the model of 'survival requires.' It means much the same as 'Morality prescribes.' Reified Morality does not require of you everything that is a necessary condition for possessing the property of morality. For example, a necessary condition for being moral

is to be alive, but Morality does not require you to be alive. Morality does not prescribe being alive.

We would not naturally say that morality requires you to be alive. This shows that in the expression 'morality requires,' 'morality' most naturally refers to reified Morality rather than unreified morality.

One consequence of reification is the one I have just described, that Morality does not require everything that is a necessary condition for being moral. But our reification of Morality goes further than that. Reified Morality also requires some things of you that are not necessary conditions for being moral. It requires some acts and omissions in the external world. For instance, Morality requires you to make sure your car's brakes are in good condition, and it requires you to refrain from murder. This is part of our commonsense understanding of Morality, and Kantians would not deny it. Being kind to strangers and refraining from murder do not supervene on your mind. Since we are making the Kantian assumption that the property of morality supervenes on the mind, it follows that they are not part of the property of morality. So on the Kantian view, reified Morality in some ways goes beyond morality as a personal property. This solves the puzzle: morality supervenes on the mind, but what Morality requires does not.

Morality is not particularly a subject for this chapter. Its relevance is that it could provide a model for the reification of rationality. We do indeed reify rationality, as I said in section 2. Let capitalized 'Rationality' be the name of the reified entity. Another name for it is 'Reason.' If Rationality is reified to the same degree as Morality, the requirements of Rationality need not supervene on the mind any more than the requirements of Morality do, even though the property of rationality does supervene on the mind.

This makes it possible for Rationality to require acts or omissions in the external world. To take one example, Rationality might require you to act only in accordance with that maxim through which you can at the same time will that it become a universal law.[1] I earlier argued that rationality cannot be identical to normativity because rationality supervenes on the mind, whereas normativity does not. But the argument does not apply to reified Rationality if it does not supervene on the mind. It could even turn out that Rationality and normativity are identical. It could turn out that what you ought to do is nothing other than what Rationality requires of you.

There would be a lot more work to do before this Kantian claim could be established. First, it needs to be explained why Rationality does not supervene on the mind. It is plausible that Morality does not supervene on the mind, because Morality is concerned with the world, not with your mind. The same cannot be said of Rationality, however reified; it is much more plausibly concerned with your mind. It is strange to claim that Rationality may require something that is not a necessary condition for being rational. Second, it needs to be explained why Rationality encompasses the whole of

[1] This is H. J. Paton's translation of Kant's categorical imperative in (Kant 1947: 88).

normativity. Even if it is granted that Rationality does not supervene on the mind, that is far short of the conclusion that it constitutes all of normativity.

Those are two big jobs to do. But if they could be done, it would mean there is a concept of Rationality in which it is identical to normativity. There would be a sense of Practical Reason that unites Rationality and normativity.

But this concept of Rationality would be very distant from rationality, the property that is possessed by people. Take a new example. Suppose Rationality requires you to act only in accordance with that maxim through which you can at the same time will that it become a universal law. I take this to mean that Rationality requires of you that you do not perform any act unless it conforms to some maxim that you can will to become a universal law. What this requirement requires of you does not supervene on your mind. For example, suppose you give money to an honest-looking person who is soliciting contributions on the street. You believe this act conforms to the maxim 'Give money to charity,' which you can will to become a universal law. Now distinguish two cases. In the first, the person is indeed collecting for charity, so your act conforms to your maxim and does not violate the requirement of Rationality. In the second case, the person is collecting for a terrorist organization, and giving money to a terrorist organization does not conform to any maxim that you can will to become a universal law. In this second case, you do violate the requirement of Rationality. Yet your mind might have all the same properties in both cases. You might never find out that in the second case you contribute to a terrorist organization; you might go to your grave believing you gave to charity. You are plainly equally rational in the two cases because your rationality supervenes on your mind. The cases differ in what reified Rationality requires of you, but your unreified rationality is the same in both.

So the Kantian project of reification does nothing to bring unreified rationality closer to normativity. It remains a mistake to identify these very different properties.

7 Summary

At the beginning of this chapter, I explained on purely semantic and grammatical grounds that 'practical reason' has two distinct meanings. It can denote the property of rationality, which people possess to varying degrees. It can alternatively denote normativity, which I also treat as a property that people possess to varying degrees – the property they possess when they do as they ought. There remained the possibility that these two properties are actually identical, but I argued they cannot be identical because rationality supervenes on the mind, whereas normativity does not.

I then warded off various responses to this argument. In a way, the most successful is a Kantian response, which does at least open up the formal possibility of a reified concept of Rationality that does not supervene on the mind. But even if the Kantian response succeeded, it would leave untouched the distinction between the property of rationality and the property of normativity. 'Practical reason' has two very different meanings, and philosophers would do well to keep them separate. This book covers two distinct topics.

Note

My thanks to Benjamin Kiesewetter and Kurt Sylvan for very helpful comments. Research for this chapter was supported by ARC Discovery Grant DP140102468.

References

Broome, John, (2013) *Rationality Through Reasoning*, Malden, MA: Wiley-Blackwell.

Fogal, Daniel, (2016) 'Reasons and reason: Count and mass', in *Weighing Reasons*, edited by Errol Lord and Barry Maguire, Oxford: Oxford University Press.

Hume, David, (1988, originally 1739–40) *A Treatise of Human Nature*, edited by L. A. Selby-Bigge, Oxford: Oxford University Press.

Kant, Immanuel, (1947) *Groundwork of the Metaphysics of Morals*, translated as *The Moral Law* by H. J. Paton, London: Hutchinson's University Library.

Kiesewetter, Benjamin, (2018) 'How reasons are sensitive to available evidence', in *Normativity: Epistemic and Practical*, edited by Conor McHugh, Jonathan Way, and Daniel Whiting, Oxford: Oxford University Press.

Lord, Errol, (2010) 'Having reasons and the factoring account', *Philosophical Studies*, 149, pp. 283–296.

Lord, Errol, (2017) 'What you're rationally required to do and what you ought to do (are the same thing!)', *Mind*, 126, pp. 1109–1154.

Lord, Errol, (2018) *The Importance of Being Rational*, Oxford: Oxford University Press.

Nagel, Thomas, (1970) *The Possibility of Altruism*, Oxford: Oxford University Press.

The Oxford English Dictionary Online, (2016) Oxford: Oxford University Press.

Prichard, H. A., (2002) 'Duty and ignorance of fact', in his *Moral Writings*, edited by Jim MacAdam, Oxford: Oxford University Press, pp. 84–101.

Ross, Jacob, (2009) 'How to be a cognitivist about practical reason', in *Oxford Studies in Metaethics, Volume 4*, edited by Russ Shafer-Landau, Oxford: Oxford University Press, pp. 243B82.

Setiya, Kieran, (2007) 'Cognitivism about instrumental reason', *Ethics*, 117, pp. 649–673.

Skorupski, John, (2010) *The Domain of Reasons*, Oxford: Oxford University Press.

Smith, Michael, (2004) 'Humean rationality', in *The Oxford Handbook of Rationality*, edited by Alfred R. Mele and Piers Rawling, Oxford: Oxford University Press, pp. 75–92.

Southwood, Nicholas, (2010) 'Vindicating the normativity of rationality', *Ethics*, 119, pp. 9–30.

Wallace, R. Jay, (2001) 'Normativity, commitment, and instrumental reason', *Philosophers' Imprint*, 1.

Wallace, R. Jay, (2020) 'Practical reason', in *Stanford Encyclopedia of Philosophy*, edited by Edward Zalta.

Wedgwood, Ralph, (2002) 'Internalism explained', *Philosophy and Phenomenological Research*, 65, pp. 349–369.

Williams, Bernard, (1981) 'Internal and external reasons', in his *Moral Luck*, Cambridge: Cambridge University Press, pp. 101–13.

Williamson, Timothy, (2000) *Knowledge and Its Limits*, Oxford: Oxford University Press.

3. Can reason be practical? Narrow and broad conceptions and capacities

Peter Railton[1]

[1] The author would like to thank the editors for exceptionally helpful comments on previous versions of this chapter. Remaining faults are my own doing.

Introduction

Can reason be practical? In order for reason to be practical, it must have a capacity to guide and motivate action, and this must be a capacity that a mind could actually put into use. But what kinds of mental processes or faculties could do this, and how? Attempts to answer these questions can take a narrower or broader form. In the narrow form, philosophers have asked about reason*ing* that is practical – self-conscious deliberation by which we can arrive at conclusions concerning what we will do or ought to do. In the broader form, philosophers have asked about how *reason* can be practical, where reason is understood as something like a capacity to be *aptly responsive to reasons for action as such*. Such a broad capacity includes the capacity for reasoning but also other capacities that enable individuals to recognize and appreciate reasons for action, including the capacity to distinguish genuine vs. merely apparent reasons and to translate reasons into action. The broader view resembles more closely the approach typical of most philosophers studying practical reason's twin, theoretical reason or epistemology. Within epistemology, self-conscious deliberation has an important role to play in understanding our responsiveness to reasons for belief, but to be rational in belief is about more than reasoning – it includes, for example, capacities for "intuition", or for non-inferential or "immediate" responses to perception, memory, or logical relations. To be rational in action likewise could involve capacities that play roles at least as central as self-conscious deliberation, and an investigation of three historically important figures in the history of thinking about practical reason – Aristotle, Hume, and Kant – will suggest that this is indeed the case. Of course, there is an element of stipulation in deciding which capacities should be grouped under the label 'practical reason'. But whatever decision is made on that score, investigating *why* a broader range of capacities might be candidates for inclusion and *how* they might actually be realized in a mind will throw into light some features of our responsiveness to reasons for action that are often overlooked. For example, it is notable that Aristotle, Hume, and (perhaps surprisingly) Kant make significant appeal to *affect* and *desire* in explaining apt responsiveness to reasons for action *as such*. We will close by asking whether the functional capacities attributed to affect and desire by these philosophers receive some support from contemporary research in psychology and neuroscience.

1 Narrow and broad conceptions of practical reason

In what we will call the *narrow* conception, practical rationality is the capacity of agents to use self-conscious, reflective deliberation to decide how one ought to act, where this deliberation is seen as a form of reasoning modeled on logical inference, leading from propositional premises to a conclusion in the form of a normative proposition about what to do:

> Statements of facts which are reasons for the performance of a certain action by a certain agent are the premises of an argument the conclusion of which is that there is a reason for the agent to perform the action or that he ought to do it... . An inference the conclusion of which is a 'There is a reason to ...' statement or an 'ought' statement is a practical inference.
>
> [Raz, 1990, 26]

Action that is "based upon" practical reasoning of this kind is typically contrasted with other, non-deliberative forms of behavior, such as reflex, instinct, or conditioned habit. Behavior resulting from non-deliberative sources need not be *unintelligent* – intelligent animals show remarkable adaptiveness in learning behaviors that meet their needs, satisfy their appetites, or promote effective social coordination across diverse contexts. But humans are thought to be not merely unusually intelligent animals but "rational animals", in Aristotle's phrase. They have a capacity to "step back" reflectively from any particular history of experience or circumstance of action, or from whatever beliefs and desires they currently happen to have, to ask whether they *should* be taking guidance from these sources. Creatures with this capacity can demand *reasons* for what they do and decide how they ought to act by deliberating over or weighing these reasons. To be sure, rational animals need not be constantly engaged in reflection of this kind – they can deliberatively formulate intentions, plans, or policies that extend over time and then rely upon these directly for guidance from moment to moment. But insofar as they are rational, such creatures can always in principle raise the question whether to continue to act upon existing intentions and plans. At least in the paradigm case, rational creatures act "under an idea" of what they are doing and why. Moreover, they can use practical reasoning to assess or revise this idea and to determine which course of action is to be taken to bring it about. This is a narrow conception of what it is for reason to be practical.

By contrast, Aristotle in *De Anima* gives a description of the "practical intellect" as beginning with motivation and ending with action: "the object of desire is the starting-point for the practical intellect", and the final step is not a judgment about how one ought to act but "the starting-point for action" itself (*DA* 433a).[1] This is a broader conception of what it is for reason to be practical: the job is not done by discovering a logical relation between reasons-statements and 'ought'-judgments; there must be motivation toward a goal, and deliberation must take us to the initiation of action to bring it about. Aristotle emphasizes that neither the starting-point of practical intellect nor the end-point of deliberation is given by reasoning alone, since *noûs* (variously translated in this context as "understanding", "intuition", or "practical insight", and contrasted with *logos* or "reason" in the narrower sense) and perception must come into play. Moreover, choice is not the same as judgment, since it combines "understanding" and "desire" to yield "deliberative appetition", which then can move us to act (*NE* 1143a, 1139b). This is a broad conception of what it is for reason to be practical: whatever constellation of capacities – perception, intuition, understanding, and deliberative appetition, as well as reasoning – is needed if reasoning is to enable us to be aptly responsive to reasons for action.

We need not decide whether the narrow or broad conception of how reason can be practical has a proprietary claim on the label 'practical reason' – perhaps it would be best to call the broader conception 'practical intelligence' or 'practical understanding'. However, all may agree that it is a desideratum of a general theory of practical reason to be able to explain how the translation from reasons for action into responses that are appropriate to those reasons is possible. Indeed, one can begin with the narrow conception and "build out" to broader conception as one assembles the pieces of such an explanation. We will begin with an inquiry into how Kant does just this.

[1] Works of Aristotle cited in the text are abbreviated '*DA*' for *De Anima* (Aristotle, 1993), '*NE*' for the *Nicomachean Ethics* (Aristotle, 1999), and '*PA*' for the *Posterior Analytics* (Aristotle, 1968). The page numbering follows standard conventions for *DA* and *NE*; for *PA* (Book.Part).

2 From a narrow to a broad conception – Kant

Unlike Aristotle, who believed that reason could not discover perfectly general rules to guide behavior in all circumstances (*NE* 1109b), Kant believed that one can "deduce" from the concept of "pure practical reason" alone an objective "supreme principle of right" entailing certain "duties" that determine right conduct. However, he recognized that this "deduction" was "analytic" and theoretical: "The supreme *principle of right* is therefore an analytic proposition" (*MM* 6:396).[1] If we are to make reason practical, it will be necessary show how rational beings can be "receptive to concepts of duty as such" in such a manner as to yield dutiful action (*MM* 6:399), and this will necessarily be a "synthetic" and practical matter. How is this to be done such that reason can "break forth into a *practical use*" (*G* 4:395) while at the same time still respecting the "purity" of the moral law? Finding an external incentive to follow the moral law might yield conduct in line with our duties, or "legality", but if we are to achieve "morality", the incentive must somehow come from the same source as the law itself (*MM* 6:218–219):

> In all lawgiving ... there are two elements: first, a law, which represents an action that is to be done as *objectively* necessary, that is, what makes the action a duty; and second, an incentive, which connects a ground for determining choice to this action *subjectively* with the representation of the law... . By the first the action is represented as a duty, and this is a merely theoretical cognition of a possible determination of choice, that is, of practical rules. By the second the obligation so to act is connected in the subject with a ground for determining choice generally.
>
> [*MM* 6:218]

The question thus becomes one of what could serve as the "*determining ground (Bestimmungsgrunden) of our will*" when our action is appropriately attuned (*Einstimmung*) to the moral law such that we act *from* duty, not simply *in accord with it*. How to interpret this attunement to the moral law is a controversial matter in Kant

[1] Works of Kant cited in the text are abbreviated as follows: '*G*' for the *Groundwork of the Metaphysics of Morals* (Kant, 1996b); *'CPrR'* for the *Critique of Practical Reason* (Kant, 1996a); '*CJ*' for the *Critique of Judgment* (Kant, 1987); '*MM*' for the *Metaphysics of Morals* (Kant, 1996c). The

scholarship, and Kant himself tells us that "how a law can be of itself and immediately a determining ground of the will (though this is what is essential in all morality) is for human reason an insoluble problem and identical with that of how a free will is possible" (*CPrR* 5:72). But Kant does suggest that we can say what the moral law "must effect ... in the mind insofar as it is an incentive" (*CPrR* 5:72), and the account he gives turns on the capacity of a recognition of value to motivate in its own right. For this to be possible, the moral law must confront us as more than a rational constraint but in a way that draws forth a motivating "positive feeling" of *respect* that has no other ground than the ground of the moral law itself: the absolute value of persons as ends-in-themselves – "*Respect* is always directed only to persons, never to things" (*CPrR* 5:76). This respect, moreover, is brought forth *immediately* by a recognition of value, without need for external incentive: "*Respect* is a tribute that we cannot refuse to pay to merit" (*CPrR* 5:77).

In order to avoid problems of regress, the capacity for respect that makes us susceptible to concepts of duty cannot be a matter of a judgment of what duty requires. It is a qualitatively different kind of state, a receptivity that can make us alive to the presence of value and gives recognition of that value practical expression in the form of action done from respect for it:

> Respect (*reverentia*) is, again, something merely subjective, a feeling of a special kind, not a judgment about an object that it would be a duty to bring about or promote. For, such a duty, regarded as a duty, could be represented to us only through the *respect* we have for it. A duty to have respect would thus amount to being put under obligation to duties.
>
> [*MM* 6:402]

We thus cannot explain the operation of respect in terms of pure practical reasoning. Instead, it depends upon a "sensibility", which is "presupposed" (*CPrR* 5:76) as part of "the mind's receptivity concepts of duty as such" that "effects in the mind" an incentive with moral force.

There is, for Kant, a close analogy with the way in which "sensibility" is presupposed as part of the mind's capacity to be "attuned" to beauty: we cannot appreciate, or be "alive to", beauty simply through the operation of reason – we must have a receptivity to beauty that is grounded in our "sensibility" and that "effects in the mind" an *appreciative* force, not a mere judgment. A creature "free from all sensibility" could not possess respect for the moral law or experience its normative force (*CPrR* 5:76). We can get a sense of respect's distinctive character by comparison with various other feelings, such *love*, *fear*, or, especially, *admiration* – attitudes that arise non-voluntarily, "as an affect", from the confrontation with a value:

---

page citations reflect the standard Akademie system: (Volume:Page).

> Something that comes nearer to this feeling [than love or fear] is *admiration*, and this as an affect, amazement, ... for example, [at] lofty mountains, the magnitude, number, and distance of the heavenly bodies [etc.].
>
> [*CPrR* 5:76]

Of course, none of these feelings *are* respect. To grasp the nature of respect as *reverentia* is to appreciate that its object must be something of absolute worth – persons (*CPrR* 5:76).

Respect, then, is the normative experience that makes us receptive to the guidance of the moral law. While it is not a voluntary attitude, "so that there can be no duty to have it", the recognition and appreciation of value it embodies leads us *freely* and *willingly* to impose the constraints of duty upon ourselves, without need for further incentive or an absurd "duty to duty". What pure practical reason *can* do is to reveal to us just how expressing this respect necessarily requires that we accept certain objective constraints with respect to others and ourselves, most notably, the constraints embodied in "the moral law". Yet, owing to our respect, these constraints are subjectively embraced as *our own*, and self-imposed by our reason:

> The consciousness of a *free* submission of the will to the law, yet as combined with an unavoidable constraint put on all inclinations though only by one's own reason, is respect for the law.
>
> [*CPrR* 5:80]

Think, by analogy, of the way in which *love* for someone involves a non-voluntary recognition of the intrinsic value of that person and at the same time an active, appreciative will to protect that person or to assist her in achieving her ends, even though this involves imposing constraints upon oneself that limit one's pursuit of one's own ends.[2] Indeed, for Kant, respect (as the positive, motivating feeling of *reverentia* and not the merely negative, restrictive "practical sense" of *observantia*, *MM* 6:449) belongs to the same family of attitudes as "love of mankind" (*Menschenliebe*):

> There are certain moral endowments such that anyone lacking them could have no duty to acquire them. – They are moral feeling, conscience, love of one's neighbor, and respect for oneself (self-esteem). There is no obligation to have these because they lie at the basis of morality, as subjective conditions of receptiveness to the concept of duty, not as objective conditions of morality. All of them are natural predispositions of the mind ... on the side of feeling.
>
> [*MM* 6:399]

[2] See, in this connection, the proposal of J. David Velleman, "Love as a Moral Emotion", that love involves a special kind of recognition of the intrinsic value of the humanity of the other (Velleman, 1999).

Like love, respect for a person both "inspires" treating that person as an end in herself and renders intelligible the constraints one thereby imposes upon one's own will (*CPrR* 5:80), constraints that rule out treating that person as a mere means to one's own ends.

It is, then, thanks to our capacity for respect as well as reasoning that the moral law can "break forth into a *practical use*" without need of any external incentive. Thus Kant is led from what seems to be a question about practical reason narrowly understood – "How can reason arrive at a supreme principle of right?" – to an answer in terms of practical reason broadly understood as incorporating both the objective determination of that supreme principle and the subjective determination of action by being appreciatively attuned to – revering – the value that stands behind it. Without such feelings as respect, an individual could theoretically cognize the analytic truth of the supreme principle of right and yet be "indifferent" to "the human being as such [as] his end" (*MM* 6:395) – and thus be "morally dead" (*MM* 6:400).

This appeal to affect or "feeling" is not a glitch. Kant's solution is perfectly precise. If agents are to respond to the moral law *for the right reason*, they must do so in recognition and appreciation of its fundamental ground: the intrinsic value of persons. When one feels love of humanity or respect for another, one is both *alive to* the intrinsic value represented by persons and *moved* to act in ways expressive of that love or respect – including the willing self-imposition of constraint that pure practical reason reveals as one's duty. Kant thus writes, "any consciousness of obligation depends upon moral feeling to make us aware of the constraint present in the thought of duty" (*MM* 6:399). But constraint is hardly the full expression of love or respect – one must also treat the other as an *end*. So "it is not enough that [the individual] is not authorized to use himself or others merely as means (since he could still be indifferent to them); it is in itself his duty to make the human being as such his end" (*MM* 6:395). Summarizing:

> The concept of duty, therefore, requires of the action *objective* accord with the law but requires of the maxim of the action *subjective* respect for the law, as the sole way of determining the will by the law. And on this rests the distinction between consciousness of having acted *in conformity with duty* and *from* duty... . It is of the greatest importance in all moral appraisals to attend with the utmost exactness to the subjective principle of all maxims ... .
>
> [*CPrR* 5:81]

Kant's solution is precise in another sense as well. Regress is avoided by means of a recognition of unconditional value, which needs no further explanation or justification but can explain and justify the self-imposition of the constraints of duty. The practical "inescapability" of obligation is thus not given by the notion of objective constraint alone but by the intrinsic value in persons that we might try to ignore but cannot wish away. Once we confront such value – and all human beings can be brought into such

a confrontation, since "No human being is entirely without moral feeling" (*MM* 6:400) – it awakens respect, as a feeling that can arise "whether we want to or not; we may indeed withhold it outwardly but still cannot help feeling it inwardly" (*CPrR* 5:77). A human lacking all such receptivity to the value of others or himself would not be recognizably human – his "humanity would dissolve (by chemical laws, as it were) into mere animality" (*MM* 6:400).[3]

But *how* does an affective state like respect motivate? All action, Kant believes, involves desire, and "The *faculty of desire* is the faculty to be, by means of one's representations, the cause of the objects of these representations" (*MM* 6:211). Affect, "feeling", functions to attach motivational interest to representations: the "capacity for having pleasure or displeasure in a representation is called *feeling*" (*MM* 6:211). In Kant's psychology, the prospect of pleasure or displeasure can generate motivation and choice: "Every determination of choice proceeds *from the representation of a possible action to* the deed through the feeling of pleasure or displeasure, taking an interest in the action or its effect" (*MM* 6:399). A "feeling", thus, can yield motivation attuned to the *evaluative representation* that constitutes the feeling: "*Moral feeling.* This is the susceptibility to feel pleasure or displeasure merely from being aware that our actions are consistent with or contrary to the law" (*MM* 6:399).

This does not mean, however, that all motivation has as its *object* pleasure or the avoidance of pain. If I desire to prove a theorem, for example, then making progress toward a proof will produce "practical pleasure", and failing to make progress will produce the "practical pain" or frustration. Thanks to the fact that my "faculty of desire" has *antecedently* set proving a theorem as a goal, an activity as intellectual as proving a theorem can have *interest* and be a source of pleasure or displeasure. But if my *end* instead were to attain pleasure or avoid pain, I would hardly be spending my hours banging my head and straining my imagination to prove a theorem.

Kant therefore distinguishes two kinds of motivation: *inclination* (*Neigung*), in which the prospect of one's own pleasure or pain is the *cause* of the interest (*CPrR* 5:73), and *desire* (*Begierde*), in which pleasure or pain is an *effect* of the interest (*MM*

[3] I should perhaps emphasize that, on the account I propose here, the "moral feeling" is no part of the *objective* ground of duty (that is, of the "*metaphysical first principles*" of duty, *MM* 6:377, which are free of any empirical determination) but rather part of *subjective receptivity* to duty. So Kant's derivation of the categorical imperative as a "touchstone" or "compass" for assessing what duty requires involves no appeal to sentiment. But this derivation is still not yet an example of, or explanation of, *practical reasoning* in the sense of reasoning that issues in motivated action – for this, some appropriate subjective receptivity to duty is needed. The moral feeling constitutes such a receptivity and is not "instinctive" or "blind", because it provides incentive to act "from being aware that our actions are consistent with or contrary to the law" (*MM* 6:399). A response that appreciates value, and is not merely a conceptual judgment that an object or action has value, can be practical *of its nature*, with an internal incentive arising immediately from its appreciative content, as a form of "attunement" to that value. For this reason, Kant analogizes the moral feeling to an affecting aesthetic response rather than a judgment: "It is difficult to think of a feeling for the sublime in nature without connecting it with a mental attunement similar to that of moral feeling" (*CJ* 5:267).

6:211). An "intellectual" or "sense-free" interest in an object can thus arise in response to an evaluative representation of an object. If my love of knowledge presents proving a theorem as in itself good or valuable, and my reason tells me that only if I impose certain constraints upon my thought will my reasoning actually constitute knowledge, then a "sense-free" interest in knowledge can become, through my faculty of desire according to concepts (in this case, proof theory), an active incentive – a potential source of practical pleasure or pain – to will self-imposition of logical constraints *from* love of knowledge, without need of external incentive and without reducing my interest to an interest of mere inclination. Similarly, when my respect for the value of persons presents them to me as ends in themselves, and my use of pure practical reason tells me that certain constraints of duty are necessary to treat others and myself as ends in themselves, then the "faculty of desire according to concepts" can give rise to an interest to will the self-imposition of those constraints upon myself *from* respect for persons. We do not have here a conflict between will and desire or between reason-based and desire-based behavior. Instead, will belongs to the "faculty of desire according to concepts", and the constraints of pure practical reason can in fact become an incentive for me thanks to my appreciation of the value that stands behind them and makes them intelligible:

> the will, as a power of desire, is one of many natural causes in the world, namely, the one that acts in accordance with concepts.
>
> [*CJ*5.172]

Because it can operate in accordance with concepts as a manifestation of respect for value, will as a power of causation is not just *causal* but *intelligible*. It places us in an order of ends, not merely causes. Thus we arrive at a "determining ground for the will" that enables us to be moved *in the right way* by the thought of duty, so that we can be alive to "moral vital force" as such (*MM* 6:400).

The ingredients of Kant's account are complex, but tracing our way through them is important if we are to understand just how delicate a task it is to show that reason can be practical. What must be shown is not simply how reasoning could issue in propositional conclusions about what duty requires. The reasoning must somehow engage *appropriate* motivation to yield *appropriate* action, where the standard of appropriateness is high: it must be a response to moral reasons in light of the kind of reasons they are – not merely as (negative) duties, but as (positive) recognition and appreciation of the value of persons.

3 Extending the broad conception – Aristotle

We have begun with Kant's philosophical approach in the realm of practical reason because it is, of all such approaches, perhaps the least likely to be suspected of being insufficiently attentive to the role of reasoning in practical rationality. But the kind of solution Kant found to the problem of explaining how reason could be practical – the way in which Kant embeds practical reasoning within a broader set of non-deliberative "faculties of the mind" that are an essential part of explaining how action could be responsive to practical reasons *as the reasons they are* – including a central reliance upon *affect* and the recognition and appreciation of *value* – is by no means peculiar to him.

If it required some effort to show how Kant's view moves *from* practical reasoning *to* the centrality of appreciation of value, in Aristotle's case, the opposite direction of movement seems to lie right on the surface – what may be of special interest, then, is to see what *kinds* of capacities he thought making reason practical involved and *how* they worked together. In particular, he located a role for reasoning *within* the scope of motivation.

Aristotle begins the *Nicomachean Ethics* with an account not of practical reasoning but of the "highest good" and unconditioned value, *eudaimonia* (*NE* 1094–1102).[1] When practical deliberation finally does receive extended discussion in its own right, in Book III, we are told that it is situated *between* ends and means and is not used to identify ends: "we deliberate about things that promote an end, not about the end" (*NE* 1112b–1113a).[2] Here Aristotle is responding in part to concerns about regress by identifying, as did Kant, a capacity for *receptivity* to value rather than an *action* such as deliberation – since action is done "for the sake of an end", and if *that* end is to be given by deliberation, we will never reach action: "if we keep on deliberating at each stage we shall go on without end" (*NE* 1113a; see also *PA* II.19).

What is this receptive capacity? Aristotle describes it as "perception" and as a non-deliberative form of "understanding" or "intuition" (*noûs*): "there is understanding, not a rational account, both about the first terms and about the last" in practical demon-

[1] Interestingly, one can argue that Kant, too, placed unconditional value conspicuously at the beginning of his best-known work, the *Groundwork*, which begins Section I with a discussion of the good will, and is clearly linked with an *appreciative* attitude: "like a jewel, it would still shine by itself, as something that has its full worth in itself" (*G* 4:394).

[2] Commentators have remarked that deliberation can also concern the *specification* of an end, in

strations (*NE* 1143a-b). To underline this point, Aristotle writes that we should pay attention to the "undemonstrated remarks and beliefs of experienced and older people or of prudent people, no less than demonstrations" since such people "see correctly because experience has given them the eye" (*NE* 1143b). Since this is a perception of an end, it has sometimes been called *evaluative perception* to distinguish it from ordinary sense perception. It is, however, engaged in sense perception – that is the idea of the "knowing eye" of those with experience and skill.

Experience in life, and skill at life, are indispensable for this kind of non-deliberative or intuitive understanding. For Aristotle, there are no general rules for conduct knowable by reason alone, and the experienced and skilled may know how to conduct themselves but not necessarily by deliberation. Full virtue or "practical wisdom" requires rational grasp of the knowledge behind the knowing eye, but ethics remains "inexact", and reason cannot supplant experience and skill. Instead, it is the widely experienced and skilled "man of good character" or "excellent person" who is able to "see what is true in every case" and who is "himself a sort of standard and measure" (*NE* 1113a).

Virtue is concerned with action, and so it requires well-developed "dispositions" or "habits" and well-attuned "feelings". For example, it is thanks to well-calibrated feelings of fear and confidence, developed through training and experience, that the brave individual is able to face danger well and to appreciate what is at risk: "the brave person's actions and feelings accord with what something is worth and follow what reason prescribes" (*NE* 1115b). Thus, "whoever stands firm against the right things and fears the right things, for the right end, in the right way, at the right time, and is correspondingly confident, is the brave person" (*NE* 1115b).

Like Kant, Aristotle thought action requires desire, and so "even if the intellect enjoins us and thought tells us to avoid or pursue something, we are not moved" until desire is brought into play (*DA* 432b-433a): "the object of desire is the starting-point for the practical intellect". Moreover, it is only in the presence of desire that deliberation can have its ending-point in "the beginning of action" (*DA* 433a). "The principle of an action", he writes, "is decision; the principle of decision is desire and goal-directed reason" (*NE* 1139a). Aristotle explains this by introducing a distinction, as did Kant, between simple appetitive motivation (*epithumia*), which takes pleasures and pains as its object, and "deliberative desire" (*boulesis*), which has as its end an apparent good and thus is susceptible to reasoning about how to obtain that good. Indeed, in practical deliberation, we deliberate *with* desire, so that decision is "deliberative appetition or appetitive deliberation", as we combine understanding with desire for an end to emerge with desire for a means – a means that might have been indifferent or aversive to us otherwise. This capacity to transfer desire from an end to an otherwise unmotivated means, even when the means is costly, difficult, and undesirable in itself, is an important mark of maturity: "this is the sort of principle that a human being is"

---

the sense of articulating a state or action that, in itself, is a realization of the end – knowledge is the end of learning, but learning is itself a realization of knowledge, not just a cause of it.

(*NE* 1139b). For genuine excellence in "practical intellect", it is not enough to bring motivation into the service of an end – the end itself must be good, and the means must be adequate to it, since only then will deliberation yield "right appetition":

> both the reasoning must be true and the desire right; and the desire must pursue the same things that the reasoning asserts. We are here speaking of intellect and truth in a practical sense: the function of practical intellect is to arrive at the truth that corresponds to right appetition.
>
> [*NE* 1139a21–28]

Aristotle's case is made easier by his view that the good for a creature is functioning in accord with its nature and that such proper functioning will yield happiness (*eudaimonia*). We therefore can become attuned through experience to our proper function by attending to the *eudaimonia* we or others do or do not experience.

Despite their many differences, there are important similarities between Aristotle's and Kant's accounts of practical reason. Both theories, in their distinctive ways, look to intrinsic, unconditional value as a fundamental determining ground of reasons for action. And neither believes that *reasoning* is sufficient to discover or appreciate this value – a *receptivity* or *sensibility* is needed, and virtue involves possession of such a receptivity or sensibility to translate value into action. For both, the "faculty of desire" must include a capacity to generate motivation in response to the recognition and appreciation of value – even if action always involves pleasure or displeasure, its incentive does not reduce to pleasure or displeasure, since we must have recourse to the valued end in order to explain how this pleasure or displeasure could even come into existence. And this capacity differentiates human action from merely animal, "appetitive", "inclination-based", or "instinctual" motivation and behavior.

To be sure, Kant was after something much more specific than Aristotle – he needed to show not only how reason could be practical but how *pure* practical reason could be practical. And Kant needed to show how this is possible without introducing the kind of natural incentive present in Aristotle's account of human function and *eudaimonia*. While both offer theories of virtue, virtue is grounded differently and plays quite different roles. And Kant's construction was correspondingly more elaborate and, one might say, more precarious. But both are applications of a framework for practical reason in which recognition and appreciation of value, in ways not accomplished by reasoning alone, play a crucial role in connecting agents with reasons, and reasons with actions.

3 Directions of fit – Hume

However much work it might take to make the case that Kant or Aristotle did not think that reasoning alone can move us to action, there is no such problem in claiming that Hume subscribed to this view. After all, Hume held that "Reason is wholly inactive" and that "An active principle can never be founded on an inactive" (*T* 3.1.1).[1] Since "*practical* philosophy" and "morals" are meant to be active, to "have an influence upon the actions and affections, it follows", he argued, "that they cannot be deriv'd from reason" alone (*T* 3.1.1).

However, some of the notions we have introduced in giving an account of Kant and Aristotle on making reason practical – evaluative representations, interests of reason, right appetition, and so on – would seem to be unavailable to Hume. Hume is thought to draw a sharp contrast between mental states that can participate in reasoning, namely *ideas* or *representations* that are capable of truth or falsity vs. *passions* or *affections* that can "influence the will" or motivate action but which are not capable of truth or falsity. He writes, in an oft-quoted passage in the *Treatise*:

> Reason is the discovery of truth or falshood. Truth or falshood consists in an agreement or disagreement either to the *real* relations of ideas, or to *real* existence and matter of fact. Whatever, therefore, is not susceptible of this agreement or disagreement, in incapable of being true or false, and can never be an object of our reason. Now, 'tis evident our passions, volitions, and actions, are not susceptible of any such agreement or disagreement; being original facts and realities, compleat in themselves... . 'Tis impossible, therefore, they can be pronounced either true or false, and be either contrary or conformable to reason.
>
> [*T* 3.1.1]

It is passages like this that have given rise to the idea of a "Humean belief-desire model of action" according to which action depends upon two fundamentally different kinds of mental states: belief-like states capable of truth and susceptible to reasoning, states defined by their "mind-to-world direction of fit"; and desire-like states capable of

[1] Herein, Hume citations beginning with '*T*' are to the *Treatise of Human Nature* (Hume, 1967) and are given in this form: (Book.Part.Section); citations to the *Abstract* are indicated as such; citations to his *Enquiry Concerning the Principles of Morals* are indicated by '*EPM*' (Hume, 1983) and are given in this form: (Section.Part).

motivating action and not susceptible to reasoning, states defined by their "world-to-mind direction of fit". No one state can, according to this model, have *both* directions of fit (Humberstone, 1992; Smith, 1987). A picture like this seems to be reflected in contemporary decision theory, which requires for decision and action both *credences* about the outcomes of possible acts and *preferences* with respect to those outcomes.

However, far from placing belief or credence on the opposite side of a divide from feelings or sentiments, Hume explicitly argued that "*belief is more properly an act of the sensitive, than of the cogitative part of our natures*" (T 1.4.1), though he recognized that this conclusion is a bit "surprizing", and he did not expect many to accept it (Hume, 1938). However, if we wish to push our investigation of the question "How can reason be practical?" to the next level, we must try to understand Hume's view of belief and why he holds it.

First, we must note that Hume's conception of *reason* in the *Treatise* is a narrow conception in the sense we have been using here – 'reason' in the *Treatise* generally refers, not to a general faculty of responsiveness to reasons, but to *processes of reasoning* or *inference*. It is notsurprising, then, that Hume joins Aristotle and Kant in thinking that "reason alone" cannot give rise to action.

Second, strictly speaking – and Hume makes it clear that he *is* speaking strictly – reasoning operates on representations, propositions, "copies", and "ideas", not upon beliefs. From 'p' and '*if p then q*', it logically follows that 'q' – the truth of the premises suffices for the truth of the conclusion. But from 'I believe that p' and 'I believe that *if p then q*', neither 'q' nor 'I believe that q' follows – the truth of the premises does not suffice for the truth of either putative "conclusion". The *object* of a belief may be a proposition with a truth value, but the belief itself is a mental state – an "original fact and reality, compleat in itself", not a "copy" of anything.

Compare the famous passage:

> A passion is an original existence, or, if you will, modification of existence, and contains not any representative quality, which renders it a copy of any other existence or modification. When I am angry, I am actually possest with the passion, and in that emotion have no more a reference to any other object, than when I am thirsty, or sick, or more than five foot high. It is impossible, therefore, that this passion can be opposed by, or be contradictory to truth and reason; since this contradiction consists in the disagreement of ideas, considered as copies, with those objects, which they represent.
>
> [T 2.3.3]

Hume, who devoted the middle book of the *Treatise* to sentiments, surely was aware that anger very often has a "reference to another object", in the sense that I can be angry *that there is no milk in the fridge*. However, Hume's point is that *anger that there is no milk in the fridge* can be factored into a "simple conception" – the idea *that*

there is no milk in the fridge – and an attitude toward it. The "simple conception" can agree or disagree with "*real* existence or matter of fact", but of itself, it contains no anger at all. The same "simple conception" appears if you are *pleased that there is no milk in the fridge.* The attitude of being pleased or angry is a "passion" with which you or I are "actually possest", not "copy" of anything, but an "original existence ... or modification of existence". So, strictly speaking – and, once again, Hume is trying to speak strictly, since he is aiming to refute a large philosophical tradition, rationalism, which he thinks owes its plausibility to a failure to attend to such matters – neither my anger nor your being pleased can be true or false, or "contradictory to truth and reason", even though their object can be.

Does this mean that Hume thought that sentiments like anger could never be more or less *apt* or *reasonable*? On the contrary, in his discussions of sentiments in Book II of the *Treatise*, Hume is interested not only their characteristic causes and effects but also in the conditions in which they are reasonably or unreasonably felt, for example, pointing out that "[someone] that has a real design of harming us, proceeding not from hatred and ill-will, but from justice and equity, draws not upon him our anger, if we be in any degree reasonable" (*T* 2.2.6) or arguing that sentiments need to be qualified by adopting other perspectives (*T* 3.2.7). But even when feeling anger or any other sentiment is reasonable, this does not mean that such a feeling is "conformable to reason" in the narrow sense – by its nature, it cannot be.

Belief, for Hume, is another such sentiment. It can take a propositional object, but the attitude of belief itself is distinct from that object or "simple conception". For example, belief *that there is milk in the fridge* and disbelief *that there is milk in the fridge* share the same object, and all the difference is to be found in the feeling toward this object with which the individual is "actually possest": "*belief is nothing but a peculiar feeling, different from the simple conception*" (*T Appendix*). Like other passions, belief is knowable by how it "feels to the mind", and forming a belief *that p* has wide-ranging effects upon our mental economy – effects that are quite different from simply contemplating the idea *that p*:

> I confess, that 'tis impossible to explain perfectly this feeling... . But its true and proper name is *belief*, which is a term that every one sufficiently understands in common life. And in philosophy we can go no farther, than assert, that it is something *felt* by the mind, which distinguishes the ideas of the judgment from the fictions of the imagination. It gives them more force and influence; makes them appear of greater importance; infixes them in the mind; and renders them the governing principles of all our actions.

/ T Appendix*]*

Like other sentiments, belief can be more or less reasonable insofar as it is proportioned to the evidence to which belief is accountable, namely evidence of truth (just as *anger* is reasonable insofar as it is proportioned to the evidence to which *it*

is accountable, namely evidence of an unjust injury). Hume provides an account of "philosophical probability", which is tied to such evidence as observed frequency of association, reflection upon parallel or analogical cases, and so on. And he goes on to say: "All these kinds of probability are received by philosophers, and allowed to be reasonable foundations of belief and opinion" (T 1.3.13). It is clear that 'reasonable' is here being used normatively – it concerns the degree of belief one *ought* to have, given one's evidence:

> Since therefore all knowledge resolves itself into probability, and becomes at last of the same nature with that evidence, which we employ in common life, we must now examine this latter species of reasoning, and see on what foundation it stands.

In every judgment, which we can form concerning probability, as well as concerning knowledge, we ought always to correct the first judgment, derived from the nature of the object, by another judgment, derived from the nature of the understanding. It is certain a man of solid sense and long experience ought to have, and usually has, a greater assurance in his opinions, than one that is foolish and ignorant, and that our sentiments have different degrees of authority, even with ourselves, in proportion to the degrees of our reason and experience.

[T 1.4.1]

Belief, then, is a "peculiar feeling", a "sentiment" or "passion", that is nonetheless assessable in terms of its responsiveness to relevant experience and reflection, that is, responsiveness to reasons for belief. Far from abolishing the Aristotelian idea that action-guiding "feelings" can be spoken of as more or less aptly responsive to reasons, Hume makes such feelings central to the architecture of his overall theory.

Including his theory of action. What, then, of desire-like states? Can they be more or less aptly responsive to reasons? Here Hume has another "surprizing" conclusion. We saw that Aristotelian desire involves a representation of its object as in some way *good*, an impression that can be mistaken, and so there is a question of when desire is "right appetition". Humean desire, by contrast, is traditionally taken to be a *non-cognitive motive force* with only "world-to-mind" direction of fit. For Aristotle (and also for Kant), acquiring a credible representation of some end as good could produce a desire to perform an action that would bring this end about, even in the absence of any pre-existing appetite or inclination toward that action. But such "desire according to concepts" seems incoherent in the Humean view. Yet Hume writes:

> The impressions, which arise from good and evil most naturally, and with the least preparation are the *direct* passions of desire and aversion, grief and joy, hope and fear, along with volition. The mind by an *original* instinct tends to unite itself with the good, and to avoid the evil, tho' they be

> conceiv'd merely in idea, and be consider'd as to exist in any future period of time.
>
> [*T* 2.3.9]

This opens the possibility of being motivated by a representation of a future good, even to perform an action to which one is now averse. Hume allows that desire can have abstract, conceptual objects that *in themselves* do not make reference to or depend upon one's own pleasure or pain, for example, "desire of punishment to our enemies, and of happiness to our friends" (*T* 2.3.9), or the desires that arise from "disinterested resentment of ... injuries to others" and "disinterested benevolence" (*EPM* 5.2, Appendix 2). Nothing in these evaluative representations depends upon our *first* finding a desire for, or aversion to, their objects; rather, it is thanks to these "disinterested" passions that we *have* a desire or aversion with respect to their objects – or for undertaking difficult or costly actions to respond to them. Hume thus characterizes the relationship of such abstract objects of desire to pleasure and pain in essentially the same way Kant would later distinguish "interests of reason" from "interests of inclination" – "These passions, properly speaking, produce good and evil [in this context, "pain and pleasure"] and proceed not from them" (*T* 2.3.9). They reverse, that is, the order of explanation from pleasure and pain to desire and aversion found in mere inclination.

For Hume, as for Aristotle and Kant, one can begin with an *evaluative appreciation* and *deliberate one's way into a desire*. Thus, if I use my understanding to discover a novel course of action that would help someone who promotes the common good, then even though this course of action involves taking steps that I find aversive, still, I can deliberate my way to being motivated to undertake these steps – even imposing upon myself standards of judgment that take me away from my personal perspective to ask what would be approved or admired from an impartial standpoint (*T* 3.2.7). As we then apply our understanding of cause-and-effect relations to the realization of these goods, "reasoning takes place to discover [these relations]; and according as our reasoning varies, our actions receive a subsequent variation" (*T* 2.3.3). To be sure, there are background passions against which such deliberation proceeds, passions that are susceptibilities on our part to various goods and evils, concrete or abstract, without which those goods would be indifferent to us. But, as we have seen, this is the same structure present in Aristotle's account of deliberative appetition and Kant's account of action in accord with duty, which look to an actually existing susceptibility to value as the subjective determining ground of decision and action.

We can also deliberate our way *out of* a desire:

> I may will the performance of certain actions as a means of obtaining any desir'd good; but as my willing of these actions is only secondary, and founded on the supposition that they are causes of the propos'd effect; as soon as I discover the falshood of that supposition, they must become indifferent to me.

[*T* 2.3.3]

This, then, opens the possibility of mistaken desire – not *false*, since desires cannot be true or false, but rather *not fitting*, because it involves a mistaken view of the nature of the object of desire. As with belief, such desire tends to go out of existence when confronted with evidence of its mistake – and that is at least one way of understanding how desire could have a "mind-to-world" as well as "world-to-mind" direction of fit. This is, of course, reminiscent of Aristotle's talk of the role of reason and understanding in arriving at "right appetition" that can possess "truth in a practical sense" (*NE* 1139a21–28).

In Hume, then, as in Kant and Aristotle, understanding the operation of practical reason involves affect as well as deliberation and involves the idea that evaluative representations, even of an abstract and "disinterested" kind, can give rise to desire. Hume adds that a similar kind of responsiveness of feeling to value exists in the epistemic case – it is thanks to a *susceptibility* to feelings of confidence or doubt in response to experience and relations of ideas that belief is possible and that casual or logical inference that take place without regress (*T* 1.3.4; *T* 1.3.7; *T* 1.4.7). And it is thanks to the capacity of affect to shape directly what we attend to, remember, expect, infer, and do, in ways that mere ideas do not, that belief can play its familiar functional role in our mental economy (*T Appendix*).

Affect has a central place in practical reason because *value* and *uncertainty* have a central place. Affect operates in the mind the way a representation of value should: it varies in character as a reflection of different kinds of potential goods, evils, or risks (e.g., fear vs. confidence, sadness vs. joy, resentment vs. guilt, etc.); it is responsive to experience and comes in degrees, which shape its action-guiding effects; it possesses positive or negative valence and can directly allocate interest and motivation; and it achieves these effects by coordinating attention, perception, inference, motivation, decision, and action. Affect moreover can both register an *appreciation* of value and shape a response appropriate to that value: fear does not only indicate risk; it presents a situation as dangerous and moves us accordingly; gratitude does not only indicate receipt of a benefit, it acknowledges the benefit and yields a favorable representation of the benefactor or beneficial act that motivates us to reciprocate; and so on. For Kant, respect is the attitude that recognizes and appreciates the intrinsic rather than merely instrumental value of persons. For Aristotle, the "doctrine of the mean" is essentially a view about the proportionality and appropriateness of certain feelings as a form of understanding of situations and and their prospects, which then can translate into appropriate action. And for Hume, belief-formation and reasoning are processes constituted by attitudes of confidence and trust, which present their objects as real or credible and shape our subsequent expectations, inferences, and actions.

Philosophers have been loath to appeal to affect in explaining our responsiveness to reasons – perhaps it would somehow appear to diminish the strictness or standing or obligatory character of our rational capacities if they involve something as subjective

as feeling. But we have seen how Kant, Aristotle, and Hume have argued that this is a mistake – without affect, even if our responses mirrored features of reasons in various ways, they would fail to recognize and appreciate these reasons and thus fail to be fully apt responses to those reasons for what they are.

4 Psychological realism?

If this understanding of Kant, Aristotle, and Hume is anywhere near the mark, then their theories of practical reason could be seen to have important implications for empirical psychology and neuroscience: we should expect to see, in the mind, capacities for what we might call *evaluative representation*, which play a central role in guiding perception, thought, feeling, and action. Moreover, we should expect those capacities to be present in the *affect and reward system*, broadly understood. Intriguingly, this is increasingly what has been found by detailed psychological and neuroscientific investigation.

Starting with the work of the cognitive social psychologist Robert Zajonc (1980), it became clear that affective responses to perceptual information appear very early in perceptual processing and shape subsequent cognition and action without need of self-conscious deliberation. Over time, a view emerged of the affective system as centrally concerned with acquiring and *appraising* information relevant to the needs, goals, and physical and social situation of individuals (Schwarz and Clore, 2003; Moors, Ellsworth, Scherer, and Frijda, 2013). Neuroscientific evidence indicated that the core structures of the affective system are first in line for the receipt of perceptual information and project widely to areas of the brain concerned with cognition, conscious experience, motivation, and action (Pessoa, 2008). The affective system moreover is a key locus of learning and memory, adapting flexibly to changing situations and forming neuronally-encoded expectations of outcomes and measures value and risk that behave in ways akin to decision theory (Preuschoff, Bossaerts, and Quartz, 2006; Moser, Kropff, and Moser, 2008; Lak, Stauffer, and Schultz, 2014). The values that appear to be encoded in the affective system cover a wide range, from basic needs to social cooperation to uncertainty, and representations of these values function as weights in choice and action (Behrens, Hunt, Woolrich, and Rushworth, 2008). Motivation is no longer seen primarily in terms of basic drives and habits but in terms of regulation by causal-evaluative models with the kinds of representational features imagined by Aristotle, Hume, and Kant in distinguishing desire from mere inclination (Berridge, 2004; Dayan and Berridge, 2014).

Of course, the state of empirical research is always complex and conflictual, and today's dominant views can become tomorrow's discarded dogmas. So we should be hesitant to infer philosophical conclusions from empirical research if we cannot find independent philosophical reasons supporting them. However, it would appear that we *can* find such reasons – most notably in the work of Aristotle, Hume, and Kant.

Centuries earlier, philosophers of practical reason "predicted" a picture of the mind that we since have seen be filled out by psychology and neuroscience.

5 Conclusion

We began with the issue of whether to conceive the question, "Can reason be practical?", narrowly (as a matter of a distinctive kind of reasoning) or broadly (as a constellation of capacities that work together to make us aptly responsive to reasons for action as such). And we have seen how three important figures in the history of practical reason – Aristotle, Hume, and Kant – recognized the need for, and filled out key elements of, a broader conception in ways that anticipated important developments in empirical psychology.

Deciding whether to use the expression 'practical reason' narrowly or broadly cannot avoid some degree of arbitrariness. It might help to avoid confusion to use the Aristotelian expression 'practical intelligence' or 'practical understanding' for the broad sense, but this would come at the cost of obscuring the essential role of the broader capacity in the operation of the narrower, privileging reason*ing* in our thinking about reason and rationality in a way that has been problematic philosophically and psychologically. In any event, what matters is that we, too, recognize the inability of the narrow conception to explain how reason can genuinely be practical, and attempt to piece together what is missing.

References

Aristotle (1993). *De Anima*, trans. by D.W. Hamlyn. Oxford: Clarendon.

Aristotle (1999). *Nicomachean Ethics*, 2nd ed., trans. by T. Irwin. Indianapolis: Hackett.

Aristotle (1968). *Posterior Analytics*, trans. by G.R.G. Mure, in R. McKeon, ed., *The Basic Works of Aristotle*. New York: Random House.

Behrens, T.E.J., L.T. Hunt, M.W. Woolrich, and M.F.S. Rushworth (2008). "Associative Learning of Social Value". *Nature 456*: 245–250.

Berridge, K.C. (2004). "Motivation Concepts in Behavioral Neuroscience". *Physiology and Behavior 81*: 179–209.

Dayan, P. and K.C. Berridge (2014). "Model-Based and Model-Free Pavlovian Reward Learning". *Cognitive and Affective Behavioral Neuroscience 14*: 473–492.

Humberstone, I.L. (1992). "Direction of Fit". *Mind 101*: 59–83.

Hume, David (1938). *An Abstract of a Book Lately Published*, ed. by J.M. Keynes and P Sraffa. Cambridge: Cambridge University Press.

Hume, David (1967). *A Treatise of Human Nature*, ed. by L.A. Selby-Bigge. Oxford: Clarendon.

Hume, David (1983). *An Enquiry Concerning the Principles of Morals*, ed. by J.B. Schneewind. Indianapolis: Hackett.

Kant, Immanuel (1987). *Critique of Judgment*, trans. by W.S. Pluhar. Indianapolis: Hackett.

Kant, Immanuel (1996a). *Critique of Practical Reason*, trans. by M.J. Gregor. Cambridge: Cambridge University Press.

Kant, Immanuel (1996b). *Groundwork of the Metaphysics of Morals*, trans. by M.J. Gregor. Cambridge: Cambridge University Press.

Kant, Immanuel (1996c). *Metaphysics of Morals*, trans. by M.J. Gregor. Cambridge: Cambridge University Press.

Lak, A., W.R. Stauffer, and W. Schultz (2014). "Dopamine Reward Prediction Error Responses Integrate Subjective Value from Different Reward Dimensions". *PNAS [Proceedings of the National Academy of Science] 111*: 2343–2348.

Moors, A., P.C. Ellsworth, K.R. Scherer, N.H. Frijda (2013). "Appraisal Theories of Emotion: State of the Art and Future Developments". *Emotion Review 5*: 119–124.

Moser, E.I., E. Kropff, and M.-B. Moser (2008). "Place Cells, Grid Cells, and the Brain's Spatial Representation System". *Annual Review of Neuroscience 31*: 69–89.

Pessoa, L. (2008). "On the Relationship between Emotion and Cognition". *Nature Reviews Neuroscience 9*: 148–158.

Preuschoff, K., P. Bossaerts, and S.R. Quartz (2006). "Neural Differentiation of Expected Reward and Risk in Human Subcortical Structures". *Neuron 51*: 381–390.

Raz, J. (1990). *Practical Reason and Norms*. London: Hutchinson, 1975; reissued by Princeton University Press.

Schwarz, N. and G.L. Clore (2003). "Mood as Information: 20 Years Later". *Psychological Inquiry 14*: 296–303.

Smith, M. (1987). "The Humean Theory of Mind". *Mind 96*: 36–61.

Velleman, J.D. (1999). "Love as a Moral Emotion". *Ethics 109*: 338–374.

Zajonc, R.B. (1980). "Preference Needs No Inference". *American Psychologist 35*: 151–175.

4. Practical reason and social practices

1 Introduction

My interest in practical reason is an outsider's interest. I am not a moral theorist or an epistemologist engaged in inquiry about reasons or rationality. In fact, I've been known to participate in good-natured ribbing about such work (but, I admit, it is a very fine line between good-natured ribbing and bad-natured ribbing, as my husband, Steve, has been known to say). But I sometimes find it useful to see my own work from an outsider's perspective: it can push me to identify and clarify assumptions that are shared by others working in the area but not outside of it; it can offer me insights into issues that I had skated over or deferred, and it can build bridges to other literatures that expand the scope of my thinking. I hope that my contribution here can do some of the same.[1]

According to a dominant model in analytic social ontology – one that plays a role in analytic philosophy from philosophy of mind, through epistemology, to ethics – the social world consists of psychologically sophisticated individuals who form intentions (consciously or unconsciously) to act. They reflect on their own beliefs, desires, or preferences, perform some sort of calculation or weighing, and, unless they suffer akrasia, perform accordingly. Sometimes they act together, or at least coordinate, under conditions of common knowledge. Sometimes they share knowledge by giving testimony or disagree with each other. Usually they say what they mean and mean what they say in a context of cooperative communication. They design practices for particular purposes and enact them for reasons. When problems arise, they must have made a factual or moral error or made a mistake in weighing their reasons. (See, e.g., Lewis 1969; Gilbert 1989; Bratman 1992.)

I don't deny that this is part of what goes on in the social world, but in order to engage in the mental activity required for this picture, we must already have quite sophisticated cognitive and linguistic capacities that include a rich supply of concepts. We must already participate in forms of interaction that enable us to make plausible interpretive hypotheses about others and that form a basis for coordination. Although some kinds of cognitive selection and patterns of interaction are hardwired in humans, there must also be forms of sociality prior to sharing intentions to take a walk (Gilbert 1989) or paint a house (Bratman 1992). Sociality does not begin with joint intentions, or the like, because our embeddedness in the social world is a precondition for even having most of our everyday intentions in the first place. These more basic forms of

[1] Throughout this essay, I draw on previous work. See Haslanger 2017, 2018 , 2019.

sociality are where we might find the sources of our practical orientations; they are the social preconditions for much of our thinking and acting.

In this chapter, I will sketch two ways in which practical reason is socially conditioned, corresponding to two ways of approaching practical reason: (i) as a human capacity for deliberation and (ii) as a normative structure. With respect to (i): most human reasoning makes use of tools that are provided by culture. The tools include language, concepts, default patterns of inference, and shared background assumptions. These tools are not simply tools for explanation and prediction, nor are they simply tools to enable us to get what we want; they are tools for coordination with others. Some of these tools are defective and leave us with unjust forms of coordination, ignorance about what's valuable, and paradoxes of agency. With respect to (ii): most coordination with others depends on there being social practices that give form to our interactions; these practices also give us reasons and shape our identities. As a result, whether something is a good or sufficient reason for action often cannot be determined without understanding the local practices that give the action meaning and play a crucial role in affecting its consequences. Tensions between practical normativity and moral normativity are well known. But the proper role and sources of social normativity – the normativity arising from participation in practices that enable and constrain social coordination – is often neglected. Of course, there is work on practical reason that fully embraces one or both of these main points (e.g., Brandom 1994, 2000; Gatens 1998; Mills 1998b; Laden 2012).[2] However, I hope that my discussion in the following can point to ways in which there are important questions about practical reason that deserve further attention and that those working in social philosophy can be part of the conversation.

[2] These references are intended as representative only, for much of feminist and critical race theory over the past four decades has explored the ways in which reasons and reasoning are socially conditioned. This is also an important theme in "Continental" philosophy since Hegel. Work in analytic moral psychology and ethics has mostly ignored these literatures.

2 Methodology, motivation, and non-ideal theory

Why do facts about human sociality matter in thinking about practical reason? Suppose we grant that practical reasoning depends on our embeddedness in social practices that frame our thinking and acting. Even so, one might argue, the basic capacity for practical reason is surely innate, and the norms of practical reason are universal. Of course, we develop abilities to reason depending on socialization; and some practices of reasoning have historical and cultural roots, for example, reasoning by reference to authority, such as the Bible or the law. But a philosophical inquiry into practical reason need not concern itself with such particulars, for the properly philosophical concern is with a more abstract form of both moral psychology and the normative structure of practical thought. Why, then, should we be concerned with the background social context?

In social/political philosophy, there has been an ongoing debate about the value of ideal theory (e.g., Mills 1997, 1998a, 2005; Sen 2006; Swift 2008; Robeyns 2008; Anderson 2010, Appiah 2017). There are many forms of this debate, but for our purposes, we can focus on two questions: (i) Should philosophical inquiry begin by considering idealized cases, selecting and abstracting from the broad range of complex and concrete phenomena which prompt our inquiry? (ii) Should philosophers aim to capture a normative ideal towards which we should be aiming (and should we undertake rectification as a process leading to that ideal)? Sometimes these two questions interact, for example, if one assumes that the ideal can be discovered only by considering idealized cases. For example, ideal theorists in political theory argue that we need to know what justice is in order to remedy current injustice, and in order to know what justice is, we must abstract away from the messy reality of our lives and understand the source and site of justice in idealized cases. (A further assumption often made is that this can be done a priori.) As Adam Swift puts it, "only by reference to philosophy – abstract, pure, context-free philosophy – can we have an adequate basis for thinking how to promote justice in our current, radically nonideal, circumstances" (2008, 382).

In recent years, the issue of ideal theory has also been raised in other philosophical domains. For example, in philosophy of language, the question arises whether we should begin theorizing with cooperative communication as the default assumption (understanding uncooperative communication in its terms, treating it as a defective case) and whether we should model ideal languages and take their features to be normative for natural language (e.g., Langton 1993; McKinney 2016). Parallel questions also arise

in epistemology (e.g., Fricker 2007; Sullivan and Tuana 2007). And the answer to such questions has increasingly been no, or not always. I myself find that the project of nonideal philosophy more effectively answers questions that I care most about. But we need not, as a discipline, make a choice; we can reasonably ask different questions calling for different methodologies. And hopefully, we can collaborate in working out the big picture.

Historically, the project of analytic philosophy has employed a method of beginning with clear and simple cases to get a grip on the phenomenon and then building up from there. Those who are dubious of this approach worry that the clear and simple cases are often not the central phenomena to be understood, and the tools we develop to understand them are inadequate and cannot simply be expanded or permuted to understand the phenomenon as a whole. A related worry is that what are in fact central, though complex, phenomena are pushed aside and downgraded as unimportant or to be addressed "later" (when we have the theory worked out!). Moreover, the selection of the clear and simple cases often seems to be biased towards cases that particular socially situated (dominant) inquirers find compelling or familiar; for example, consider the invisibility of care work in most political philosophy. And finally, it seems methodologically questionable to develop theory using mainly a priori methods and then think it can be simply applied as needed, for example, as if so-called "applied" ethics can be done by just adding empirical details to utilitarianism or deontology.

Inquiry begins with a question. Of course, there are many kinds of legitimate questions. But one set of legitimate questions concerns how we might solve a specific problem facing us, a problem that has arisen in our interaction with the world, with each other, or through self-reflection. To solve such problems, idealized examples and ideal theory are neither necessary nor particularly helpful. In particular, we need to understand the complexity of the particular situation that faces us to diagnose the problem (also drawing on empirical inquiry); we can make progress in solving the problem without knowing what the ideal is and attempting to implement an ideal when the circumstances and agents are far from ideal is often a mistake.

One source of the problem with ideal theory is that when we think of humans "in the abstract," it is tempting to begin with a single person. They have a body and a mind and are capable of using the mind to move the body to perform actions. Of course, humans live with others, and an individual's action affects the lives of others. So we should be attentive to this interaction in understanding and evaluating individual actions in relation to others. But this abstraction is inadequate, at least for some purposes. For example, if our guiding question concerns the terms on which we should organize life together, and if we begin with the singleton agent with capacities for deliberative thought and agency (including motor skills), then very young children, some of the disabled and elderly, and non-human animals are left as afterthoughts (Nussbaum 2006). If the bodies and the histories of the abstract agents are washed out as well, then reproductive and sexual differences, historical injustices, and material conditions are occluded. This is not to say that we cannot learn from philosophical

exercises that focus on idealized abstract humans. But for many of us, it is at the very least unclear how to apply those lessons to the problems that motivate our inquiry, for we are not trying to understand how abstract humans might organize themselves. We want to solve problems that arise in organizing our lives together here and now.

Critical social theory is a form of nonideal theory that begins inquiry with problems arising in nonideal circumstances and in resistance to injustice. For example, one pressing concern is why even thoughtful and well-meaning people act in ways that contribute to injustice – and do so in ways that perpetuate their own subordination or the subordination of those they love. Surely, most of us are not knowingly and intentionally dominating others or allowing ourselves to be dominated. Yet this happens nonetheless. A rather straightforward example is the division of labor in the household, that is, women's "second shift" (Hochschild and Machung 2003). Even those who are conscientiously egalitarian and even feminist in their politics live in ways that burden women with housework, childcare, eldercare, and care of the sick and disabled that far exceed their fair share. Why do we continue to do this? We are not ignorant or morally corrupt, at least not in any straightforward sense. In some accounts of practical reason, it is not obvious how to analyze or evaluate such action. Are oppressed individuals not being appropriately responsive to reasons? How should we weigh the reasons one has to participate in unjust systems of coordination, especially if an unjust system is the only one available? What are we to make of desires – and conceptions of the good – that are formed under such conditions?

One way of capturing this problem is in terms of *ideology*. Stuart Hall suggests that ideology

> has especially to do with the concepts and the languages of practical thought which stabilize a particular form of power and domination; or which reconcile and accommodate the mass of the people to their subordinate place in the social formation.
>
> (1996/2006, 24–25)

The first challenge of a theory of ideology is to understand how we, collectively and voluntarily, enact social structures. The more specific, and more pressing, question is how, without being coerced, we come to enact oppressive social structures. Of course, not all oppression or injustice is ideological: "a whole number of other factors ... can play an important role ... , from selectively applied repression via coordination and cooperation problems in the face of massive power asymmetries to the 'pathologies' and paradoxes of collective action" (Celikakes 2016, 20). Ideology, nevertheless, is a kind of barrier to social justice, one that ultimately affects our agency "from within."[1] It is tempting to say that ideology creates an epistemic barrier, but it is more than

[1] For more on the construction of "docile" subjects (though not employing the terminology of "ideology"), see Foucault 1979; Bartky 1990.

that, for it affects not only our perception and belief formation but a wide range of affective, conative, and hedonic states and processes and bodily dispositions (Railton 2014).

I will use the term "practical orientation" for the shared (or coordinated), often unconscious, broadly psychological dispositions that enable us to engage with the world (including other agents) around us. In my view, an ideology is, as Hall suggests, a set of social meanings – public symbols, scripts, and other cultural tools – that we internalize and use to frame our thought and action and, moreover, systematically sustains injustice.[2] An individual's practical orientation is ideological to the extent that it is shaped by cultural tools that – in a particular context – produce or sustain injustice. Such practical orientations will involve a kind of reason-responsiveness (Mantel 2018; Lord 2018). But because our practical orientations are shaped by ideology and under conditions of injustice, they are liable to structural distortion. The critical theorist, then, begins with practical questions such as: What parts of my practical orientation (and the orientations of others) can I trust? How do structures of injustice colonize our thoughtful and well-intentioned engagements with each other and the world? The starting point of inquiry is not an abstract agent but, rather, individuals with minds and bodies that have been shaped by interactions with others and whose actions are meaningful primarily within social practices.

In the next section, I will sketch an approach to culture that illuminates how culture might play an essential role in our practical orientations. I will then turn to consider briefly how social practices draw on culture to enable us to coordinate, and in doing so, give us reasons to act. I am not arguing, however, that all reasons are dependent on practices or that reasons are necessarily constituted within practices. In my view, practices not only provide reasons but also occlude them. Ideology distorts our practical orientation both by shaping the possibilities of coordination on unjust terms and also by *preventing* us from recognizing what is morally valuable and imagining coordination on better terms. Some reasons are not (easily) epistemically accessible from within our practices; it is only through challenges to the practice that we gain access to them. I will conclude by making some of the connections between these ideas and practical reason more explicit.

[2] Theorists use the term "ideology" in many different ways. There is, for example, both a pejorative and a non-pejorative sense (Geuss 1981). In a non-pejorative sense, ideology guides our participation in social practices, whether just or unjust. In the pejorative sense, the term is used as part of an explanation of how unjust and oppressive social structures are stabilized and sustained: it is an attempt to illuminate how our agency has been colonized. It may be that there is not just one kind of thing going wrong in the variety of different cases. But the hypothesis is that there is a meaningful difference between practical orientations that systematically sustain injustice and those that don't, and the former are ideological. In recent years (since ~2015), I use the term in the pejorative sense.

3 Culture: giving shape and content to our thinking

The term 'culture' has been highly contested in the social sciences and humanities for decades. What counts as culture, or a culture, is not only a descriptively challenging question but is also normatively laden. Is there a meaningful notion of culture that we can draw on in thinking about practical reason?

One concern is that there are plausibly two different notions of culture, and eliding them has politically problematic effects (Appiah 2016). In the *Tylorian* (Tylor 1871) conception, culture is "that complex whole which includes knowledge, belief, arts, morals, law, customs, and any other capabilities and habits acquired by man (sic) as a member of society" (Tylor, quoted in Appiah 2016, 2). In the *Arnoldian* conception, culture is*:* "a moral and aesthetic ideal, which found expression in art and literature and music and philosophy" (Appiah 2016, 2; see also Arnold 1869/2006). The Arnoldian conception focuses on what is sometimes called "high culture," as opposed to "popular culture." When we speak of culture in this sense, we typically use the singular: culture is what artists and humanists, as opposed to engineers, builders, or inventors, create. If we assume that to have a culture is to have the "high" culture of (European) elites, then the rest of us are downgraded to (more or less) barbarians.

Another concern with the concept of *culture* is that its employment too easily gives rise to a kind of cultural relativism. In the Tylorian conception, we are members of a society by virtue of internalizing its culture. In a strong version of the Tylorian view, we become who we are by virtue of internalizing the norms and values of our society's culture, and we cannot truly understand the norms and values of a culture that we haven't internalized. Values and norms can be appreciated only "from the inside," that is, from practitioners. From this, two conclusions seem to follow: (i) we cannot fully understand members of other cultures, and (ii) critique misfires, since there is no neutral standard – no values or set of norms – that can provide a basis for critique. One need not accept this strong (one might say, hegemonic) version of Tylorianism. Societies are fragmented and have multiple, often conflicting, sets of social meanings; navigating the (contemporary) social world involves code-switching and shaping.

Contemporary social theorists reject both the Tylorian and Arnoldian conceptions of culture and rely instead on a much more fragmented, pragmatic, polyvocal, and creolized conception of culture. Begin with the idea that any human behavior is conditioned by multiple factors. Suppose we are hungry and look for food. We might ask different questions about such a sequence of behavior, and find different factors

relevant, for example, the *physical* demands of the human body, the *geographical* context and the edible things in it, the *social/political context* that makes certain edibles salient and available, the *economic constraints* on what the individual(s) can afford, or the *social meaning* of the different foodstuffs: do we go for a burger and French fries or Buddha's delight? Within any such sequence of behavior, physical and cultural processes interact with each other and "[such] interaction is only one of the many ways in which the cultural forms part of and is continuous with the natural world" (Balkin 1998, 5).

Social meanings are captured and expressed in language, but not just in language. We recognize and respond to a broad range of symbols, signs, statuses, and so on and navigate the world with default assumptions (what some might consider analytic or quasi-analytic truths) that shape our engagement with each other and the non-human world. Understanding semiotic relations as relations in a holistic web of meanings, William Sewell (2005) suggests:

> culture is not a coherent system of symbols and meanings but a diverse collection of "tools" that, as the metaphor indicates, are to be understood as means for the performance of action. Because these tools are discrete, local, and intended for specific purposes, they can be deployed as explanatory variables in a way that culture conceived as a translocal, generalized system of meanings cannot.
>
> (46)
>
> It is important to note that the network of semiotic relations that make up culture is not isomorphic with the network of economic, political, geographical, social, or demographic relations that make up what we usually call a "society." A given symbol [e.g.,] mother, red, polyester, liberty, wage labor, or dirt, is likely to show up not only in many different locations in a particular institutional domain (motherhood in millions of families) but in a variety of different institutional domains as well (welfare mothers as a potent political symbol, the mother tongue in linguistic quarrels, the Mother of God in the Catholic Church).
>
> (49)

So in this Sewellian account, culture is a set of tools that human and some non-human agents employ in thinking and acting (see also Balkin 1998, Ch. 1, Lessig 1995). Some tools are simple meanings (pink means girl, red means stop); some are narrative tropes ("First comes love, then comes marriage, then comes baby in the baby carriage"); some are default assumptions ("Marriage is between one man and one woman") or heuristics (imitate-the-majority or imitate-the-successful (Hertwig et al 2013, 7; Gigerenzer et al 1999); some are familiar patterns of metaphor and metonymy ("Juliet is the sun," "The pen is mightier than the sword," Camp 2006); some are

entrenched conceptual homologies (reason : passion :: man : woman) (Balkin 1998, Ch. 10; Balkin 1990). These tools form complex but fragmented and negotiated networks of meaning. I share some cultural tools with philosophical colleagues across the globe, others with my neighbors, and still others with my dogs. The public meanings are internalized as we learn the "languages" of our local cultures; this is the basis for our fluency in social skills and capacities for code-switching. Furthermore, there are forms of thinking, feeling, and acting that do not require culture – some individuals cannot access most cultural tools due to their cognitive differences, and sometimes we can rely on what Grice called "natural" meanings. But culture is important, one might even argue essential, for broad human coordination.

If we accept the idea of culture as a set of tools, then some of the risks of the Tylorian and Arnoldian conceptions are diminished. Culture is not a unified and coherent system; we engage in different practices with different communities (at work, at home, at leisure, in religious communities) that require different semiotic tools and different background assumptions. Tools that are designed for one purpose are used for other purposes and can be used for coordination or critique. Moreover, one's culture doesn't "determine" what one does (in a way that compromises autonomy) any more than one's language determines what propositions one expresses. But culture provides substantial content for our mindedness, and its contribution is aptly subject to critique, for this is one site where ideology can infiltrate our lives by shaping not only beliefs (and patterns of ignorance) but the full range of human responsiveness, including desires, emotions, preferences, perception, habits, dispositions, and the like.[1] Insofar as practical reason guides us in navigating the social realm and coordinating with others, we cannot avoid drawing on the resources – good or bad – that culture provides.

[1] Fricker (2007) has drawn attention to a related issue of *hermeneutical injustice*. In Fricker's view, however, hermeneutical injustice is treated as a harm to an individual by virtue of a lack of a hermeneutical resource. The phenomenon I have in mind is more structural and productive. Our capacity for social agency presupposes, as Rawls (1955) would say, a "stage setting for action." What I can do and who I can be depend on the communicative and interpretive resources that culture provides.

4 Social practices

In the previous section, I argued that culture shapes our responsiveness to things; it does so by providing us a set of tools for interpreting and engaging the world and by creating a set of (fragmented, dynamic, context-sensitive) frameworks of intelligibility. It is crucial for these tools and frameworks to be public and shared, for they form the basis for coordination. Coordination around *resources*, that is, things of (+/-) value, is a fundamental human task, and our ability to develop flexible forms of coordination that can be passed down through social learning is the key to our evolutionary success (Sterelny 2012). Coordination is embodied in social practices. In the account I favor, social practices are patterns of learned behavior that, at least in the primary instances, enable us to coordinate as members of a group in creating, distributing, managing, maintaining, and eliminating a resource (or multiple resources), due to mutual responsiveness to each other's behavior and the resource(s) in question, as interpreted through shared meanings/cultural schemas.[1] The cultural schemas/meanings are the tools provided by culture that I discussed in the previous section.

In the philosophical literature on practices, there is a common assumption that social practices are rule-governed patterns of behavior. Although I do not share this assumption (in my view, only some practices are, strictly speaking, rule governed), insights from this literature are valuable for elaborating the social embeddedness of practical reason. In "Two Concepts of Rules" (1955), Rawls argues that practices (such as promising) are defined by a set of rules that are *logically prior to* the behavior and states of mind of the participants. The practices render our action meaningful. Standard examples of this include games: one cannot score a soccer goal without there being a set of rules that constitute soccer and an occasion in which the game is being played (and performing certain behavior counts as a play in the game, whether one intended it to or not). Practices also constitute reasons for action: Jozy Altidore has reason to pass the ball to a teammate or into the goal and not touch the ball with his hands, because this is what soccer requires. Commitment to the game, and to his team, has broader consequences in his life and gives him reason to practice, to travel to certain destinations, to manage his diet and other activities (Chang 2013). Such actions may be constitutive of his identity, who he is.[2]

[1] I explicate and defend this view in Haslanger 2018. The view, as I understand it, presumes value pluralism (see Anderson 1993). Note that I am concerned with *social* practices; in principle, a pattern in an individual's behavior may be a practice but not a social practice.

[2] Games are not the only, or even the best, examples. Not all practices are constituted by rules;

Once we recognize that at least some reasons are practice dependent, there will be many cases in which we cannot evaluate an agent's reason for action without understanding the social context.[3] As Tamar Schapiro has pointed out,

> Because the actions falling under practice rules are logically constituted by those rules, such actions can only be justified by being shown to be in accordance with those rules. To justify a move by showing its conformity to some standard that is independent of the game is to justify it as some other, practice-independent form of behavior, rather than as the move that it is.
>
> (2003, 335)

Of course, the practice itself may call for justification, and such justification may rely on practice-independent standards such as utility. (And evaluating practices is itself a practice!) However, as long as one remains a participant in the practice – and for many practices and roles, simply choosing to opt out is not an option – the rules of the practice are a source of reasons, and we can do better or worse, with more or less justification, even in deeply problematic practices. How do we adjudicate the weight of practice-dependent reasons and the weight of reasons for the practice? For example, if we simply assume that reasons for or against *A-ing* are given by the constitutive standards of the practice within which *A-ing* occurs, then it is not clear how to accommodate systematic critique.[4]

For example, in the context of social roles, we are often put in a position of a forced choice. As a professor, you must assign your students a grade. The grading system is given. You can decide on the method and standards for assigning grades, but you are not in control of what grades are available or their meanings. You can choose not to assign a grade, but that too has a meaning and consequences for the student. In a forced choice, both action and inaction have significance, so there is no way to avoid a move in the practice. In choosing how to act, the structure of the practice shapes my reasons. I give Genae an A because those who perform excellently in all components of evaluation should be assigned an A, and she met this standard. This is the grade I ought to give her.

---

we do not always engage in all practices consciously and deliberatively, and what practice my action instantiates is not just up to me or my intentions. (See Rebecca Kukla and Marc Lance 2014.)

[3] Recall that I am not arguing that all reasons are practice dependent. For example, practices can shape our practical orientations so that it is difficult, if not impossible, to recognize reasons to oppose the practice or to do things very differently.

[4] For example, consider Schroeder (2010, 13): "According to this account, the distinction between the right and wrong kind of reasons is relative to an 'activity'. This is because the point of the distinction between the 'right' and 'wrong' kinds of reasons, is that only the 'right' kind contribute to standards of correctness, and standards of correctness are relative to activities." Thanks to Kurt Sylvan for directing me to Schroeder's paper.

Social practices, however, interact in complicated ways. Recently I was at an institute-wide curriculum retreat in which we discussed grading systems. It was pointed out that in most science classes at MIT, an A grade is rare and is reserved for only exceptional students (and there are no + or − options, so the next grade below A is a B). However, to get into a good medical school, students need to have an A average in their science classes; a B average is insufficient. So the MIT grading system effectively prevents many very qualified and capable MIT students from going to medical school. This prompted a valuable discussion about the purpose of grades and how the multiple purposes might be best achieved. What practice would be best? And what, under the current system, should a responsible professor do? Would a professor who gives every passing student an A be fulfilling their responsibilities? As you might expect, there was much disagreement and debate and no resolution.

This would appear to be a case of overlapping practices: there is a local practice of grading that enables faculty at MIT to coordinate, to motivate their students, and to treat them fairly. There is a broader practice of using grades for assessment that extends to other professional contexts. There are different goals amongst the institutions, different practical orientations amongst the students, teachers, and administrators. And there is a background (though contested) assumption of meritocracy. Note, however, that in order for any of the broad goals to be achieved, a practice must be put in place that will enable assessment and sharing of relevant information. The idea that there is a correct grade for a student, apart from a set of practices that interprets and employs the information conventionally encoded, is like saying that there is a correct word for a two-wheeled vehicle, apart from any language. In effect, there is no reason for someone to assign an A, a B, or a C without a set of practices and conventions that use the grade assignment to enable coordination. There are many acceptable ways to set up such practices. And even if the form of coordination is less than ideal, a participant in the practice will have reason to conform to it, because it may be the only means of coordination available.

Schapiro explores the form that practices take under nonideal conditions. There are several ways that conditions may be nonideal in a cooperative practice. Consider a practice involving the coordination of two individuals (she uses the practice of negotiation as a paradigm). (i) One's co-participant may sincerely participate in the practice but frustrate its ends by failing to do their part effectively or responsibly, for example, they may do it too slowly, carelessly, without being fully aware of their role, and so on. Schapiro would call one's action in such a case *productively* unsuccessful (337). (ii) One's co-participant can fail to meet the constitutive conditions of the practice, for example, they may not undertake it sincerely or may lie about having met the preconditions. This would be a case where one's action is *constitutively* unsuccessful. Due to the misdeed(s) of the other, the practice, Schapiro suggests, becomes a sham. (iii) One's co-participant may *subvert* the practice, making it the case that by playing one's part, one does not contribute to achieving its ends but to something at odds with its ends; perhaps one contributes, unwillingly, to their misguided ends. In such a case,

one's participation conforms to the rules but is at odds with the spirit of the practice, but the failure occurs by virtue of another's misdeed. (iv) One's co-participant can exploit your commitment to the practice for their own ends: even if all participants are committed to the end of the practice, your participation may have side effects that another takes advantage of; for example, your teammate may want to win the game as much as you do but also take advantage of your exhaustion after a hard practice (when your willpower is low) to borrow money. The practice becomes a method for the co-participant to get you to do what he wants (345).

As Schapiro points out, however, nonideal conditions are not just a matter of having an uncooperative partner in negotiation; the structure of rules and circumstances can effectively undermine one's agency in the same sorts of ways. For example, consider a double-bind:

> because you are a participant [in the practice], you have to play, and because the rules are constitutive of participation, the only way to play is to play by the rules. But because the background conditions presupposed by such rules are ill established, playing by the rules fails to amount to participation in the relevant form of activity. As such, what is a conscientious player to do? If you comply with the letter of the law, you will betray its spirit, in the sense that you will not be engaging in the form of activity in terms of which you value yourself and your conduct as a player. If you violate the letter of that law, however, you will likewise fail to participate in that form of activity, because there are no other rules in terms of which that activity is defined.
>
> (Schapiro 2003, 340)

Think back to the grading problem. A professor has to assign a grade, and to do so, must follow the rules. But the rules are not well suited to achieve communication and coordination (given the gap between what an A means in different contexts); any strategy of assigning grades within the current system will succeed along one dimension but will fail along another (and so betray the spirit[s] of the practice). But making up my own grading system will not work, because it won't constitute a basis for coordination on assessment – my idiosyncratic system of assigning say, Q, L, H, does not define a new social practice. The possibility of scenarios such as those listed previously makes clear that we are not simply dependent on others to be efficient or productive in satisfying our desires or preferences but that the constitutive possibility of action can depend not only on performing according to practice rules but also on cooperative others and on the background social conditions. We might say that social normativity is a kind of normativity that derives from practices due to their (broadly) conventional means of facilitating (but not guaranteeing) coordination.[5]

[5] In my view, not all conventions are conventions in Lewis's (1969) sense, and not all coordination

In considering Rawls and Schapiro, I've assumed that practices are rule governed and that we participate with others intentionally. Let's go back, however, to the idea that an individual's practical orientation is shaped by culture and that their social fluency in cooperative practices happens, for the most part, without thought or deliberation. Culture is not a set of rules (those who lack social skills are often excellent at following rules – knowing or acting on rules is not the problem). And culture shapes the content of our mental lives, from perception, to intention, to inference. Let's grant, further, that in nonideal circumstances, culture engages us, systematically, in unjust systems; that is, it is an ideology. We desire fashionable things at the cheapest price that are often produced by exploiting labor and the environment; we defer to the powerful; we avoid the discomfort of diverging from social scripts and are anxious around (sometimes violent towards) those who seem to have different scripts. We try to follow the "letter of the law," but the law isn't written and we have to do the best we can with ambiguities and vagueness; even the spirit of the law is unclear; for example, people not only wonder what etiquette requires but even what the point of etiquette is (cf. Judith Martin 2005).

Under such nonideal conditions, looking for reasons for action or justifying actions one has performed is a messy business. In some cases, it is difficult even to determine what action one has performed. Have the background conditions been met? Are others participating sincerely? Have they subverted the practice? Should you subvert the practice towards a better end? Is social engineering required (changing the practice), or is reform possible (bringing the practice in line with its proper end) (Schapiro 2003, 352)? I don't raise these question to answer them but to complicate the examples that could be the subject matter for philosophical reflection.

is a solution to a formal coordination problem. The solutions may not be arbitrary; there may not be, in any meaningful sense, common knowledge among participants; the responses may not be rational or mutually advantageous. I am dubious, especially, that preferences should be our starting point. A meaningful sense of preferences with respect to the resource in question may be constituted only through the practice that organizes our responses (Anderson 2001).

5 How, again, is this relevant to the study of practical reason?

I've sketched a case for thinking that most agency, including rational agency, depends on being embedded in a culture, for culture provides basic tools for mindedness. As mentioned previously, I am not immersed in the study of practical reason. However, as I understand the subject, a better understanding of the social context of agency might have something to contribute. For example, we might add complexity to some of the core questions as follows:

i) What *are* reasons? How can we gain knowledge of reasons for action?

I have not provided an account of reasons. Some normative reasons are practice dependent, for example, I have reason to give my student an A for her excellent performance in class. But under nonideal conditions, there may also be reasons to resist or reform the practice that do not derive from that practice or even other existing practices. The tools that our existing cultures provide for coordination are often defective: they deflect our attention from things that are morally and practically relevant, for example, our interdependence with other species in an ecosystem; they distort our ability to detect or create value (think of the fetishism of the commodity). Practices we rely on for coordination impose social hierarchy (Tilly 1999). Practices not only create, shape, and make reasons vivid, they occlude them, prevent us from acting on them, and usurp our good intentions. We have some reason to participate in existing practices, even bad ones, for opting out can be costly or impossible under conditions of forced choice. I would hope that studies of practical reason could help us sort through the challenges of living within unjust practices and illuminate the structural liability to structural dysfunction they impose.

How do we gain knowledge of reasons for action? Some knowledge we gain by becoming fluent participants in the practice. But, as noted, this is not good enough. Fortunately we are participants in many practices and can gain critical perspective on one practice from engaging in others. One might find reasons to resist traditional gendered practices at home by entering differently gendered practices in the workplace; open-minded travel to other cultures (which may be just "across the tracks") is also important. Although I have not argued for it here, I believe that we can also have first-person affective knowledge of reasons to resist domination.

> ii) What capacities are central to being a rational agent? What capacities are central to being a moral agent? What is the relationship between deliberation and (moral, practical, etc.) action?

Many of us are capable of deliberation about how to act, but our participation in everyday practices is mostly routine, and there is little resistance. We develop practical orientations – complex cognitive, affective, and agential dispositions – to respond to each other and the world around us. These practical orientations are deeply social: our thinking, feeling, and acting are structured to coordinate with others in a particular milieu that has been shaped to facilitate that coordination.[1] We are not and have never been isolated individuals just trying to make it in the world.

Crucial to the success of our practical orientations is a complex set of social meanings that are publicly recognized and, if necessary, enforced. We have gendered pronouns, and gender is considered a deep and important fact about each of us. But the gender of an infant is not obvious. So we use colors to code infants' dress, bedding, toys, and other equipment; names and other bodily styling (such as earrings) also provide signals. The information such coding supplies enables us to integrate children "properly" into our gendered practices. We are fluent in reading the gendered signals (deliberation is not required) and respond with differential treatment. Failing to do so is considered rude, sometimes offensive. I would argue that fluent participation in social practices of this sort is a form of rational agency, even if we are not conscious of the signals we are responding to and even if the meaning of our action in a particular context is not (fully) under our control (Lessig 1995; Haslanger 2018). But the fact that we participate as rational agents in such practices does not render them (or us) immune from critique.

Work on practical reason might fruitfully engage with problems that emerge when our practical orientations are ideologically distorted. Under nonideal conditions, we are often just doing our best to coordinate with others on terms that they can interpret as meaningful. And this seems reasonable. But even if we become aware problems and distortions, often we cannot simply refuse to play. (We have to choose some pronoun, some words, some responses.) Yet, for many of us, we go against deep commitments in doing so. It is hard to discern what would be rational in such circumstances; or if rationality is a matter of acting to maximally satisfy our (socially conditioned) preferences, it is hard to see why rationality is a virtue.

> iii) What is the normative structure of practical reason? How *ought* we reason when deliberating about how to act? What (perhaps retroactively or from an objective point of view) justifies an action? What makes an action reasonable?

I have not even sketched how one might develop a normative account of practical reason; instead I simply raised a series of questions that arise when agency is embedded

[1] On the importance of social niche construction, see Mameli 2004; Sterelny 2012; Zawidzki 2013.

in unjust social structures. I have assumed that social practices can provide reasons. I am obliged to keep (most of) my promises, because that's what the practice of promise-keeping requires. I am obliged to thank my hosts for a lovely dinner and suggest that we must have them to our house soon, because that's what the practice of gift-giving (around here) requires. One might argue, however, that practices provide *normative* reasons only if they are themselves well formed and warranted; and some practices only *seem to* provide us with normative reasons. But what does that evaluation entail? Our practices have complex histories, and society does not provide us with a unified and coherent set of rules. Against what social and empirical background should we undertake the evaluation of a practice? Moreover, such a demand is overly restrictive. There are better and worse ways of going on, even when our systems of coordination are not well established or create injustice.

For example, refusing common courtesy is not a transgressive act of courage; it is rude and disrespectful. As Judith Martin (N.d.) states:

> serving as the language and currency of civility, etiquette reduces those inevitable frictions of everyday life that, unchecked, are increasingly erupting into the outbursts of private and public violence so readily evident in fractured families, stymied legislatures, drop-of-the-hat lawsuits, road rage, and other unwelcome by-products of a manners-free existence. These unpleasant developments have bred a nationwide call – from academics, politicians, writers of all stripes, and the public at large – for a return to common courtesy.

But participating in common courtesy may be a process of self-subordination, given the unjust systems some forms of courtesy are designed to protect. Ideological oppression exploits our motivation to engage in cooperative practices and depends on such double binds. Under such conditions, we may need to subvert the practices or at the very least reform them. I would hope that research on practical reason could illuminate our agency within nonideal circumstances and offer tools to help us responsibly navigate the unjust and dysfunctional structures that we embody. Social norms are not arbitrary or optional, for they provide the structure of life together, and we can't simply opt out of our form of life. But neither should we simply conform to their demands, for to do so is to become complicit in injustice.

Acknowledgments

Thanks to Åsa Burman, Ruth Chang, Katharine Jenkins, Mari Mikkola, Tamar Schapiro, Kurt Sylvan, Aness Webster, Charlotte Witt, and Stephen Yablo for helpful conversation and feedback.

References

Anderson, Elizabeth. (1993). *Value in Ethics and Economics*. Cambridge, MA: Harvard University Press.

———. (2001). "Unstrapping the Straightjacket of 'Preference': A Comment on Amartya Sen's Contributions of Philosophy and Economics." *Economics and Philosophy* 17: 21–38.

———. (2010). *The Imperative of Integration*. Cambridge, MA: Harvard University Press.

Appiah, K. Anthony. (2016). "Mistaken Identities: Creed, Country, Color, Culture." The Reith Lectures, Lecture 4: Culture. BBC Radio 4.

———. (2017). *As If: Idealization and Ideals*. Cambridge, MA: Harvard University Press.

Arnold, Matthew. (1869/2006). *Culture and Anarchy*. Oxford: Oxford University Press.

Balkin, J.M. (1990). "Nested Oppositions." *Yale Law Journal* 99(7): 1669–1705.

———. (1998). *Cultural Software: A Theory of Ideology*. New Haven: Yale University Press.

Bartky, Sandra Lee. (1990). "Foucault, Femininity, and the Modernization of Patriarchal Power." In *Femininity and Domination*. New York: Routledge, pp. 63–82.

Brandom, Robert. (1994). Making It Explicit: Reasoning, Representing, and Discursive Commitment*. Cambridge, MA: Harvard University Press.*

———. (2000). *Articulating Reasons: An Introduction to Inferentialism*. Cambridge, MA: Harvard University Press.

Bratman, Michael. (1992). "Shared Cooperative Activity." *The Philosophical Review* 101(2): 327–341.

Camp, Elisabeth. (2006). "Metaphor and That Certain 'Je Ne Sais Quoi'." *Philosophical Studies* 129: 1–25.

Celikates, Robin. (2016). "Beyond the Critical Theorists' Nightmare: Epistemic Injustice, Looping Effects, and Ideology Critique." Presentation at the Workshop for Gender and Philosophy, MIT. May 12.

Chang, Ruth. (2013). "Commitments, Reasons, and the Will." In Russell Shafer-Landau, ed., *Oxford Studies in Metaethics*, vol. 8. Oxford: Oxford University Press, pp. 74–113.

Foucault, Michel. (1979). "Docile Bodies." In Alan Sheridan, trans., *Discipline and Punish*. New York: Vintage Books, pp. 135–169.

Fricker, Miranda. (2007). *Epistemic Injustice*. Oxford: Oxford University Press.
Gatens, Moira. (1998). "Institutions, Embodiment, and Sexual Difference." In Moira Gatens and Alison Mackinnon, eds., *Gender and Institutions: Welfare, Work and Citizenship*. Cambridge: Cambridge University Press.
Geuss, Raymond. (1981). *The Idea of a Critical Theory: Habermas and the Frankfurt School*. Cambridge: Cambridge University Press.
Gigerenzer, Gerd, Peter M. Todd, and The ABC Research Group. (1999). *Simple Heuristics That Make Us Smart*. Oxford: Oxford University Press.
Gilbert, Margaret. (1989). *On Social Facts*. Princeton: Princeton University Press.
Hall, Stuart. (1996/2006). "The Problem of Ideology." In Kuan-Hsing Chen and David Morley, eds., *Stuart Hall: Critical Dialogues in Cultural Studies*. New York: Routledge, pp. 24–45.
Haslanger, Sally. (2017). "Culture and Critique." *Proceedings of the Aristotelian Society, Supplementary Volume* 91: 149–173.
———. (2018). "What Is a Social Practice?" *Royal Institute of Philosophy Supplements* 82.
———. (2019). "Cognition as a Social Skill." *Australasian Philosophical Review* 3(1): 5–25.
Hertwig, Ralph, Ulrich Hoffrage, and The ABC Research Group. (2013). *Simple Heuristics in a Social World*. Oxford: Oxford University Press.
Hochschild, Arlie, and Anne Machung. (2003). *The Second Shift*, 2nd edition. London: Penguin Books.
Kukla, Rebecca, and Mark Lance. (2014). "Intersubjectivity and Receptive Experience." *The Southern Journal of Philosophy* 52: 22–42.
Laden, Anthony Simon. (2012). *Reasoning: A Social Picture*. Oxford: Oxford University Press.
Langton, Rae. (1993). "Speech Acts and Unspeakable Acts." *Philosophy and Public Affairs* 22(4): 293–330.
Lessig, Lawrence. (1995). "The Regulation of Social Meaning." *University of Chicago Law Review* 62(3): 943–1045.
Lewis, David. (1969). *Convention: A Philosophical Study*. Oxford: Blackwell.
Lord, Errol. (2018). *The Importance of Being Rational*. Oxford: Oxford University Press.
Mameli, M. (2004). "Nongenetic Selection and Nongenetic Inheritance." *British Journal for the Philosophy of Science* 55(1): 35–71.
Mantel, Susanne. (2018). Determined by Reasons: A Competence Account of Acting for a Normative Reason. *New York: Routledge.*
Martin, Judith. (2005). *Miss Manners: Guide to Excruciatingly Correct Behavior*, freshly updated edition. New York: W.W. Norton & Co.
———. (N.d.). "About Miss Manners." www.missmanners.com/about-miss-manners/.
McKinney, Rachel. (2016). "Extracted Speech." *Social Theory and Practice* 42(2): 258–284.

Mills, Charles. (1997). *The Racial Contract*. Ithaca: Cornell University Press.
———. (1998a). *Blackness Visible: Essays on Philosophy and Race*. Ithaca: Cornell University Press.
———. (1998b). "White Right: The Idea of a *Herrenvolk* Ethics." In Charles Mills, ed., *Blackness Visible: Essays on Philosophy and Race*. Ithaca: Cornell University Press, pp. 139–166.
———. (2005). "'Ideal Theory' as Ideology." *Hypatia* 20(3): 165–183.
Nussbaum, Martha. (2006). *Frontiers of Justice: Disability, Nationality, Species Membership*. Cambridge, MA: Harvard University Press.
Railton, Peter. (2014). "The Affective Dog and Its Rational Tale: Intuition and Attunement." *Ethics* 124(4): 813–859.
Rawls, John. (1955). "Two Concepts of Rules." *Philosophical Review* 64(1): 3–32.
Robeyns, Ingrid. (2008). "Ideal Theory in Theory and Practice." *Social Theory and Practice* 34(3): 1–62.
Sen, Amartya. (2006). "What Do We Want from a Theory of Justice?" *The Journal of Philosophy* 103: 215–238.
Sewell Jr., William. (2005). "The Concept(s) of Culture." In Gabrielle M. Spiegel, ed., *Practicing History: New Directions in Historical Writing after the Linguistic Turn*. New York: Routledge.
Schapiro, Tamar. (2003). "Compliance, Complicity and the Nature of Nonideal Conditions." *Journal of Philosophy* 100(7): 329–355.
Schroeder, Mark. (2010). "Value and the Right Kind of Reason." *Oxford Studies in Metaethics* 5: 25–55. Page number cited is to the manuscript available here: https://philpapers.org/archive/SCHVAT-2.pdf.
Sterelny, Kim. (2012). *The Evolved Apprentice*. Cambridge, MA: MIT Press.
Sullivan, Shannon, and Nancy Tuana, eds. (2007). *Race and Epistemologies of Ignorance*. Albany: SUNY Press.
Swift, Adam. (2008). "The Value of Philosophy in Nonideal Circumstances." *Social Theory and Practice* 34(3): 363–387.
Tilly, Charles. (1999). *Durable Inequality*. Berkeley: University of California Press.
Tylor, Edward Burnett. (1871). Primitive Culture: Research into the Development of Mythology, Philosophy, Religion, Art, and Custom*, Volume 1. London: John Murray.*
Zawidzki, Tadeusz W. (2013). Mindshaping: A New Framework for Understanding Human Social Cognition*. Cambridge, MA: MIT Press.*

5. How to be a pragmatist

1 Pragmatism and the limits of normative intuitions

Pragmatism is said to be the view that we should "do what works." This seems like empty and useless advice, since it omits any substantive criterion of what works. Indeed, pragmatists reject the quest for a fundamental principle of right action, or a definitive criterion of the good, as misguided. Instead, they advise us to replace this quest with *methods* for intelligently updating our moral beliefs.

This advice does not leave us rudderless. After all, in the empirical sciences, we also lack substantive criteria of cognitive success that lie outside of scientific practice.[1] All we have are the methods developed by science itself. Pragmatists hold that normative inquiry is of a piece with any other kind of empirical inquiry. Whatever standards of success exist in normative inquiry are internal to normative practices. This does not commit us to relativistic complacency. Just as the empirical sciences regularly improve their methods in light of experience, so can normative inquiry improve its methods.

We can sharpen this idea by contrasting dominant methods of moral philosophy with pragmatist methods. Both take practical intuitions – feelings or ideas about what one ought to do, how one ought to feel, or what is a reason for or against doing or feeling something – as indispensable materials for practical reflection. But dominant methods of moral philosophy suppose that such intuitions can deliver far more than what pragmatist philosophers think is credible. Dominant methods suppose that we can use intuitions and a priori reasoning to discover fundamental principles of morality that are systematic – that ground and unify our deliberations across all or at least large domains of conduct. They presuppose a sharp separation of non-normative facts from ultimate values, such that, with sufficient reflection, we can arrive at fact-free foundational normative principles that are true or applicable in all possible worlds.[2] Dominant methods also pay limited and relatively superficial attention to problems of

[1] Nor should we jump to the conclusion that truth is the only or ultimate cognitive aim of science. Many scientific theories sacrifice accuracy for simplicity, ease of cognition, computational tractability, or susceptibility to empirical testing. Idealizations, which by definition are not strictly true, are ubiquitous in science.

[2] See Cohen, G. A. 2003. "Facts and Principles." *Philosophy and Public Affairs* 31(3): 211–45 for a contemporary defense of this view. Kant, Immanuel. [1785] 1981. *Grounding for the Metaphysics of Morals*. Trans. James Ellington. Indianapolis: Hackett offers a classic defense. Some advocates of armchair intuitive methods deny that they yield fundamental moral principles, however. See, for example, Dancy, Jonathan. 2004. *Ethics Without Principles*. New York: Oxford University Press.

moral bias in our intuitions.[3] They may pay some consideration to the possibility that our intuitions are biased by self-interest. Thus, we may be asked to reflect from behind a veil of ignorance, in which we don't know certain particular facts about ourselves or to consider that we may be the recipient rather than the doer of some action we are inclined to perform, as a check against self-interest. But dominant methods largely neglect other biases that can affect moral reasoning – for example, biases that inform the very ways we frame the normative questions to which we seek answers, which may neglect the concerns of those not consulted in normative inquiry.

Pragmatist methods have more modest objectives. They do not seek principles that purport to be true in all possible worlds. Rather, they seek no more generality in normative principles than is required for the resolution of moral problems that we are (or realistically risk) confronting.[4] Pragmatist methods, relieved of the demand to seek necessary principles, are free to make use of empirical methods for updating our moral beliefs. They pay greater attention to many types of psychological and social causes of bias, making full use of empirical discoveries in the social sciences about how people reason under different conditions. They also stress that practical principles themselves are tested empirically, in practice.

A naturalistic perspective on normative intuitions supports the modest pragmatist view. In the standard case, normative intuitions arise in deliberation. Deliberation is a kind of thought experiment undertaken by an agent who encounters a need to act but uncertainty about how to proceed. To resolve this uncertainty, the agent entertains

[3] Some philosophical approaches take arguments concerning bias in intuitions to the opposite extreme. Utilitarians have used them to discredit deontological moral intuitions altogether in favor of the principle of utility. See Bentham, Jeremy. 1907. *An Introduction to the Principles of Morals and Legislation.* Oxford: Clarendon Press, ch. 2, for a classic example. I have argued that utilitarians, in practice, cannot credibly dispense with deontological intuitions in calculating utilities (Anderson, Elizabeth. 1993. *Value in Ethics and Economics.* Cambridge, MA: Harvard University Press, ch. 4). Recently, experimental philosophers have cast doubt on the normative force of intuitions by appealing to interpersonal variation in normative intuitions. See, for example, Weinberg, Jonathan, Shaun Nichols, and Stephen Stich. 2001. "Normativity and Epistemic Intuitions." *Philosophical Topics* 29(1–2): 429–60. Such arguments can lead in radically relativist or subjectivist directions. From a pragmatist point of view, the difficulty with all such positions is that they regard intuitions in categorical and static ways. Since, in any naturalistic account of normative judgment, we can't do without intuitions, the task before us is rather to consider how our intuitions can be improved, how we can form better intuitions through bias-reducing methods. While experimental philosophy may prove useful for testing such methods, it has not been systematically deployed to that end.

[4] For example, in Anderson, Elizabeth. 2010. *The Imperative of Integration.* Princeton: Princeton University Press, ch. 7–8, I supply a contextualized rationale for antidiscrimination principles that explains why they are consistent with affirmative action programs. Many philosophers attempt to justify antidiscrimination principles by seeking some key feature that makes an act of discrimination, considered in isolation, inherently wrong. This quest makes affirmative action programs appear to be inherently unjust. Against such approaches, I argue that certain antidiscrimination norms of *limited* scope are justified as useful tools for dismantling existing systematic group-based disadvantage. Certain types of affirmative action program are likewise justified as tools for counteracting or weakening our implicit group biases.

various courses of action, imagines the expected consequences of those actions, and simulates her normative reactions to these actions and their consequences. Normative intuitions just are these simulated reactions. In deliberation, the agent undertakes this thought experiment with the intention to act as her normative intuitions direct.

When philosophers elicit normative intuitions in thought experiments such as the famous trolley cases, these intuitions are typically the conclusions of *simulated deliberation*.[5] Philosophers forecast the consequences of various actions and simulate their valuations of them. However, in simulated deliberation, we elicit intuitions *without* any intention to act as they direct. We are thinking without having any real stakes in the outcome. These thought processes are less serious, in a practical sense, than planning. Moreover, they are further removed from experience than ordinary deliberation. To ensure that principles arrived at could be true in all possible worlds, the dominant philosophical methodology dictates that we entertain thought experiments about bizarre cases distant from prior experience and work up principles that can encompass all our intuitions for all cases. The hope is that we could thereby arrive at foundational normative principles once and for all, which would then be applicable in all future cases.

No one supposes that actual deliberation leads us to non-trivial necessary truths. Often enough, even when we act on the most thoughtful deliberation, we end up doing something we regret. Intuitions elicited in philosophical thought experiments – in *simulated* deliberation – can hardly be more reliable than actual deliberation. It is subject to the same errors and biases. Like the conclusions of actual deliberation, the conclusions of simulated deliberation tend to be less reliable, the further removed they are from circumstances with which the thinker is familiar.

Consider Judith Jarvis Thomson's famous thought experiment in which she asks us to consider a society in which its members reproduce by means of floating seeds that embed themselves in the carpet.[6] If someone carefully screens their windows to prevent the entry of "people seeds" but some manage to get in anyway, would they be obligated to cultivate the seeds? Thomson's thought experiment aims to elicit intuitions that are supposed to be equally applicable to the case of a woman in our society who conscientiously uses birth control that happens to fail. Should she be denied the right to abortion because her birth control failed? On the assumption that intuitions give us insight into moral principles that are true in all possible worlds, this bizarre thought

[5] One might also take up the perspective of the recipient of action or of a detached moral observer in the thought experiment. Each perspective – first-person, second-person, and third-person – yields distinctive insights into moral dimensions of the action. Dewey, John. 1925–1953. "Three Independent Factors in Morals." In *The Later Works*, ed. Jo Ann Boydston. Carbondale, IL: Southern Illinois University Press, 279–89 argues that these perspectives yield insights into the good, the right, and the virtuous, respectively; that each is an independent and irreducible element in determining what one ought to do; and that there is no definitive, context-free formula for resolving tensions in these perspectives. For simplicity of exposition in the main text, I focus on the first-person (deliberative) perspective. But my remarks on its limitations apply as much to the other perspectives.

[6] Thomson, Judith Jarvis. 1971. "A Defense of Abortion." *Philosophy and Public Affairs* 1: 47–66.

experiment would be a reasonable way to figure out what those principles would be. But consider the fact that any society whose members reproduced as imagined would be organized in radically different ways from ours in innumerable other ways not captured in Thomson's thought experiment. There is no reason to suppose that we have any reliable access to the moral norms that would make sense for beings who reproduce that way. We should not give much or any credence to intuitions about such bizarre cases.

This gives us reason to think that the conclusions philosophers draw from bizarre thought experiments are not trustworthy. Even in more ordinary cases, intuitions elicited in philosophical thought experiments are probably *less* reliable than deliberation, because they are less serious and more removed from experience. No extra credit should be given to them because they are arrived at in leisure. Planning, which is a type of deliberation, can also be done at leisure. Nor is it necessarily better that philosophical thought experiments are undertaken in the "cool hour," without personal stakes. The weakening of emotional cues and personal stakes while simulating deliberation may well lead us to overlook the importance of certain reasons and to speculate in irresponsible ways, as it may lead us to neglect considerations that are rightly colored by strong emotions.

Yet we cannot do without normative intuitions. Instead of burdening them with unreasonably high expectations for what they can deliver, pragmatists urge us to abandon the quest for necessary normative truths. Instead, we should search for methods for intelligently updating our current intuitions or normative judgments to address the problems we currently face and similar problems we realistically risk facing.

This raises the question of how to judge the reliability of our normative intuitions. Consider how, in actual practice, we do so. Suppose someone deliberates about a practical problem and acts on the conclusion of her deliberation. She will then confront the *actual* consequences of her action and the *actual* normative reactions of herself *and others* to her act and its consequences. Sometimes, these consequences, or normative reactions to them, are both surprising and negative. This is the occasion for reactive emotions such as regret, remorse, shame, and anger directed at the act and its agent. Such reactions, *when they are surprising*, are signs that the agent made a mistake in deliberation, that her original intuitions were erroneous.[7] She then may have reason to revisit her deliberations to see where she went wrong.

[7] If these consequences and reactions to them were anticipated, although the agent followed the conclusions of her deliberation, they need not signal any error in the agent's original deliberations. Sometimes, people know they have to make hard decisions that will elicit negative reactions from others. Sometimes, people know they are stuck with habitual negative reactions to certain acts, although they reject the normative reasoning that could rationalize such reactions. Consider the lapsed Catholic who rejects the church's teachings against divorce but who still feels lingering guilt about having divorced her abusive husband. She may rationally dismiss her guilty feelings as the product of indoctrination into false views. This is *not* to claim that reflective judgments are always sufficient grounds for dismissing anticipated negative reactions. If someone still can't get over such reactions even after a long time,

she may not have gotten to the bottom of her normative concerns and may have reason to reopen reflection on the matter. Nor do I claim that surprising reactions are always reliable guides to normative considerations. They are *evidence* of the presence of values or reasons. Evidence is always defeasible.

2 Pragmatist methods as problem solving

The key to understanding pragmatist methods for improving our normative judgments is to consider the instrumental value of value judgments. We *use* them to guide our conduct and valuations of things. We formulate them in order to *solve practical problems*: to figure out what we should do when we are uncertain about how to proceed but need to act; to change the ways we value certain things when our current valuations have gotten us into trouble; or to figure out how to value a novel object. Because we use value judgments to solve practical problems, such judgments are subject to empirical testing: we can see whether these judgments actually help us solve the problems we are using them to solve. Practical judgments or policies imply hypotheticals of (roughly) the form: "adhering to principle x (in conduct or deliberation) will help solve or ameliorate problem y."[1] If such hypotheticals fail when tested in their expected contexts, we can investigate what went wrong with our judgment-forming process to see whether it involved systematic flaws, the correction of which would enable us to generate more reliable judgments.[2]

[1] See Dewey, John. 1976. "Valuation and Experimental Knowledge." In *The Middle Works, 1899–1924. Vol. 13, 1921–1922*. Carbondale, IL: Southern Illinois University Press, 3–28. We may also consider evaluative judgments that point us to (un)desirable outcomes. Such judgments imply hypotheticals of (roughly) the form: "if you experience x, you will (dis)like it." They are useful for reminding us why we should not give into temptation or why some daunting effort toward some remote end is worthwhile and thereby help us overcome problems of temptation and intimidation. Once we view value judgments as problem-solving tools, we have further reason to doubt the aspiration of dominant methods of normative philosophy, to discover foundational normative truths that hold in all possible worlds. Why think there is any tool that could solve all normative problems (or even all problems of a specific type, such as moral problems) in all possible worlds?

[2] Philosophers often drive a sharp wedge between the truth and the usefulness of value judgments in a way that would discredit the claims of pragmatist methods to yield normative knowledge. For example, Derek Parfit argues that the true normative theory might be "self-effacing" or such that it ought not be believed (Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Clarendon Press, 23–4). I argue against this view in Anderson, Elizabeth S. 1991. "John Stuart Mill and Experiments in Living." *Ethics* 102: 4–26. Of course, pragmatism recognizes that a judgment may expresses an *ideal* that would be found worthwhile were it to be achieved, even though *intending* to follow it is self-defeating. For example, a color-blind society might be a valuable ideal, even though adopting color-blind *principles* of deliberation (never permitting conscious considerations of race in the distribution of goods) in our world reinforces race-based disadvantage by barring remedies needed to counteract and weaken the operation of implicit, unconscious racial biases. See Anderson, Elizabeth. 2010. *The Imperative of Integration*. Princeton: Princeton University Press, ch. 8.

This general account suggests two types of methods for improving our value judgments. First, we can test our judgments through *experiments in living*. In contrast with the dominant methods, which test judgments only in thought experiments, pragmatists urge us to learn from the ways our uses of such judgments fare in practice. We test our judgments by living in accordance with them and seeing whether they solve the problem we are trying to solve with acceptable side-effects. Conclusions drawn from actual experiments in living are more reliable than those drawn from deliberation or simulated deliberation in philosophical thought experiments, because they are based on more accurate information – about the actual consequences of actions and our actual normative reactions to those actions and their consequences. *Normative claims, just like causal claims, are subject to experimental testing.*

Second, we can adopt *bias correction methods*. When we find that we have acted on a bad judgment – one that fails to solve our problem with acceptable consequences – we can investigate the processes that generated that judgment to determine where we made a mistake. This is something that we do already after having made a stupid mistake: we go back to our decision-making process and ask where we went wrong, with the aim of revising that process in such a way as to avoid that mistake in the future. Pragmatists recommend that we take advantage of scientific research to aid error detection and improve our decision-making processes. In many cases, we discover that our deliberations were biased in a systematic way. Research in the cognitive sciences, social psychology, and other branches of the social sciences can help us detect and diagnose systematic biases in our thinking and design practices or institutions to correct or counteract those biases. These bias-correction methods are themselves subject to empirical testing to determine whether they are actually effective in correcting or counteracting the biases in question.

Too see how these methods work, we need to consider the occasions in which the need for them arises. Most of the time, we act out of habit, including habits formed on the basis of prior value judgments and social norms, which are institutionalized rules of conduct, usually expressing value judgments, that are reciprocally followed by many people. Sometimes, however, we encounter situations in which the habit does not offer clear guidance or indicates a course of conduct that is impossible in the circumstances. Sometimes we find that acting on the habit generates new or newly recognized bad consequences. Call such occasions *problematic situations or experiences*. They force us to stop automatically behaving and reflect on what to do. They expose a practical defect in our prior habits and hence in the value judgments or other mental states that they incorporated. To solve our problem, we need to devise a new value judgment that can correct or counteract or work around those defects. In other cases, we acted with full deliberation, but the judgment on which we acted failed to solve our problem or generated surprisingly bad consequences or surprisingly negative reactions to anticipated consequences. Such circumstances also put us into a problematic situation, requiring us to make a new judgment to guide action and perhaps also inspiring us to reflect on

what went wrong with our original judgment so we can improve our decision-making processes.

Updating our value judgments can be schematized as a multistep process. In this discussion, I will focus on cases of *moral* failure – failure with respect to what we owe to each other.

The first step is to survey the problematic features of the situation that prompted reconsideration of the initial intuition. In the case of moral failure, this typically involves unresolved interpersonal conflicts or negative reactions of other people to the agent's act. There is now uncertainty or disagreement on how to proceed in light of the surprising negative consequences of habits or judgments that seemed to be working until now.

Second, we need to provisionally diagnose the problem that now needs to be solved. In cases of moral failure, diagnosis usually needs to focus on the negative experiences and complaints of those who are suffering or discontented with the agent's action or policy. In contrast with philosophical thought experiments, here the agent needs to attend to the actual problems of actual people rather than the hypothetical problems of imaginary people in bizarre situations.

Third, we should reflect on the deficiencies of current habits, norms, and normative judgments for coping with the problem as diagnosed. It could be, for example, that current judgments were designed to cope with different problems or conditions or crafted without awareness or appreciation of all the consequences of acting on them. It could be that the normative concepts habitually applied fail to discriminate features of the situation that are now salient parts of the problem, suggesting the need for conceptual revision.

Fourth, we may consider whether the defects in the original judgment or practices based on it reflect systematic biases that led to error. For example, social psychologists inform us that ethnocentrism, or ingroup favoritism, is a pervasive moral bias.[3] If complaints about an agent's problematic conduct are originating from groups to which the agent does not belong, it may be that ethnocentric bias led the agent to unjustly neglect the interests or perspectives of outgroups.

Fifth, such reflections may suggest that we institute bias-correcting or bias-blocking procedures or institutions based on empirical evidence of what kind of practices effectively counteract such biases. The normative intuitions, judgments, or evaluations resulting from such corrective procedures are likely to be more reliable than before. A model of how this works can be found in the institution of double-blind placebo-controlled trials for judging the effectiveness of medical treatments. Such procedures are designed to block wishful thinking on the part of patients and doctors so that more accurate judgments of medical effectiveness can be made.

[3] Donald Kinder and Cindy Kam provide an excellent overview and synthesis of theories of ethnocentrism in Kinder, Donald, and Cindy Kam. 2009. *Us Against Them: Ethnocentric Foundations of American Opinion.* Chicago: University of Chicago Press, ch. 1.

Sixth, we test our new practical judgments by acting in accordance with them and evaluating the results. Two questions are relevant in evaluation. First, does acting on the new judgments solve the problem as originally diagnosed? And second, does it do so with acceptable side effects? An affirmative answer to both questions amounts to a successful test of the new judgment in an experiment in living. Usually, however, some unanticipated negative consequences arise from acting on the revised judgment. That is, the experienced valuation of outcomes (our normative reaction to them in experience) differs from our original ex ante evaluation of outcomes. If they are bad enough, this may prompt a new diagnosis of the problem. Alternatively, it may prompt a revision of evaluative criteria or desiderata that incorporate attention to the negative consequences. We may repeat our process until we arrive at value judgments, action on which solves our problem as diagnosed with consequences we feel we can live with.

In describing the several steps of pragmatist reasoning, I use the plural "we" deliberately. Moral reasoning – reasoning about what we owe to each other – is fundamentally joint or collective, because it is essentially about making demands, requests, or other kinds of claims on others, which purport to give reasons to those others to respond accordingly.[4] Those others are expected to respond to the reasons proffered but may reject them, make contrary claims based on other reasons, and so forth.

In fact, much normative reasoning, even about the good, relies on collective reasoning. For example, insofar as individuals seek excellence in some domain of achievement, such as athletics or scholarship, they participate in social practices with collectively devised standards of value. They may propose revisions of those standards. A gymnast may invent a new type of move that spurs a revision in scoring. A scientist may criticize a particular statistical technique for analyzing data and propose an alternative that avoids the defects of the original. These are proposals for revision of shared standards, not simply an expression of idiosyncratic tastes. Meaningful achievement is something that can be recognized as such by others.

[4] Darwall, Stephen. 2006. *The Second-Person Standpoint: Morality, Respect, and Accountability.* Cambridge, MA: Harvard University Press.

3 An example: recognition of slavery as a moral wrong

John Dewey's naturalistic social psychology argues that humans are already equipped to follow pragmatist methods and indeed often already do so implicitly.[1] The key to improvement is to do so knowingly, with the full assistance of the empirical sciences to help us determine the consequences of our experiments in living, potentially problematic biases of reasoning, and corrective methods *and* with the full engagement of others in discussion to diagnose problems and forge, test, and revise solutions to problems we face together – a practice he called democracy.[2] *Continuous improvement of the methods of science and democracy are the means by which we improve our practical reasoning.*

All of this may seem either too abstract or too anodyne to be helpful. In any event, for a pragmatist, the proof of any theory of practical reason is in the pudding: how well does it actually work in practice? To see how pragmatist methods work, we need to see them in action. I shall briefly illustrate these methods by pointing to episodes in what may be the most consequential transformation of collective moral consciousness in modern times: the worldwide change of belief concerning the moral status of slavery. Three hundred years ago, almost no one believed that chattel slavery was morally wrong. Over the course of the 19th century, anti-slavery convictions triumphed across the entire Western hemisphere and Europe. By the mid-20th century, they encompassed the globe. This came in conjunction with the worldwide abolition of chattel slavery as a legal institution.[3] Today, almost no one is willing to argue for its reinstatement, and its wrongness is taken for granted in legal and public moral discourse.

The full story of this stunning moral transformation is far too complex to cover in this brief article, particularly since it was achieved by different means in different countries. One point is clear, however: the change in moral convictions was not produced

[1] Dewey, John. 1922. *Human Nature and Conduct: An Introduction to Social Psychology.* New York: H. Holt and Company.

[2] Dewey, John. 1927. *The Public and Its Problems.* New York: H. Holt and Company; Dewey, John. 1988. "Creative Democracy: The Task Before Us." In *The Later Works, 1925–1953, Vol. 14: 1939–1941, Essays*, ed. JoAnn Boydston. Carbondale, IL: Southern Illinois University Press, 224–30.

[3] I choose my words carefully. Involuntary servitude in many forms, such as debt peonage and penal labor, is still legal in many parts of the world. Trafficking is still widely practiced, although it is illegal. What has been abolished worldwide (with the possible exception of Mauritania) is chattel slavery, understood as the consignment of human beings to the legal status of moveable property in

by pure moral argument alone. To be sure, abolitionists tried to persuade by moral arguments, hoping to convert slaveholders one by one to the conviction that slavery is wrong and from there to freeing their slaves. Yet pure moral arguments addressed to slaveholders mostly fell on deaf ears. Proslavery writers met each antislavery argument with mountains of rebuttals.[4]

Moral psychology helps us understand why moral argument alone was ineffective. It wasn't simply that slaveholders' material self-interest got in the way. Many non-slaveholding whites, too, who had no immediate material stakes in slavery, saw nothing wrong with the practice. This is connected to the racialized character of slavery in the West, which put whites on a higher social stratum than blacks due to their immunity from slavery, in contrast with blacks' liability to that degraded status. Standing in a position of superior power over others tends to bias the moral sentiments of the powerful, in at least three ways: it reduces their compassion, activates their arrogance, and leads them to objectify subordinates. Call these effects "power biases."

Psychologists have shown that higher-class individuals are less compassionate than lower-class individuals.[5] It is not surprising, therefore, that proslavery writers confidently argued that slaves did not suffer from the cruelties of slavery. For example, William Harper, South Carolina State Representative and prominent advocate of slavery, insisted that whipping "is not degrading to a slave, nor is it felt to be so."[6] Senator James Henry Hammond of South Carolina claimed that slaves by nature did not suffer from forced separation from their relatives when families were broken up by sales to distant slaveholders.[7] Such views were characteristic of proslavery writers, who regularly claimed that slavery was beneficial to the slaves.

John Dewey and James Tufts, in their famous ethics textbook, argued that power tends to make people arrogantly confuse their personal desires with morality: "It is difficult for a person in a place of authoritative power to avoid supposing that what he wants is right as long as he has power to enforce his demand."[8] Individuals learn the difference between what they want and what is right through experiences in which others hold them to account for wrongdoing through practices such as blame and pun-

---

another's possession.

[4] See, for example, Elliott, E. N., ed. 1860. *Cotton Is King, and Pro-Slavery Arguments: Comprising the Writings of Hammond, Harper, Christy, String Fellow, Hodge, Bledsoe, and Cartwright, on This Important Subject.* Augusta, GA: Pritchard, Abbott & Loomis.

[5] Piff, Paul K., et al. 2010. "Having Less, Giving More: The Influence of Social Class on Prosocial Behavior." *Journal of Personality and Social Psychology* 99(5): 771–84; Stellar, Jennifer E., Vida M. Manzo, Michael W. Kraus, and Dacher Keltner. 2012. "Compassion and Class: Socioeconomic Factors Predict Compassionate Responding." *Emotion* 12: 449–59.

[6] Harper, William. 1860. "Slavery in Light of Social Ethics." In *Cotton Is King, and Pro-Slavery Arguments*, ed. E. N. Elliott. Augusta, GA: Pritchard, Abbott & Loomis, 575–6.

[7] Hammond, James. 1852. "Letters on Slavery." In *The Proslavery Argument as Maintained by the Most Distinguished Writers.* Charleston: Walker, Richards & Co., 132.

[8] Dewey, John, and James Tufts. [1932] 1981. *Ethics, The Later Works, 1925–1953.* Ed. Jo Ann Boydston. Carbondale, IL: Southern Illinois University Press, 226.

ishment. To the extent that the powerful escape such accountability in some domain of conduct, they are liable to be morally confused in that domain. Slaveholders, exercising arbitrary and unaccountable power over slaves, and whites generally holding power over blacks (since blacks in the United States and Britain could not sue whites in court, even if they were free), therefore had an impaired ability to distinguish their desires from their moral duties to slaves and to blacks generally.

The racialized practice of slavery induced a third bias in whites: objectification. A dominant social group objectifies a subordinate group when it *views* the latter in terms of its service to the dominant group's desires, *enforces* that view by placing the latter in a subordinate role in service to those desires, and *represents* subordinates as inherently fit for that role and unfit for superior roles.[9] Objectification is an instance of what psychologists call the "fundamental attribution error," whereby observers misattribute a person's behavior to their innate characteristics rather than the situation in which they are placed.[10] It occurs when the observers are also the agents who constrain the observed to behave in the ways attributed to their natures. Through slavery, whites enforced their view that blacks exist to serve whites' needs, while proslavery ideology represented blacks as inherently fit for slavery. James Hammond expressed objectification bias in his notorious "mud-sill" speech before the Senate in 1858, in which he argued that all societies needed a class of people consigned to menial work, and that the South consigned blacks to such work because they were fit by nature for slavery.[11]

This partial account of power biases – a subset of the moral biases implicated in the practice of slavery – helps us see why moral argument alone was insufficient to change people's minds. Arguments do not dislodge the circumstances that induce moral bias and thereby make it difficult for people to grasp the moral objections to slavery and act accordingly.[12] This account also points to measures that could be taken to counteract these biases. *The statement of the problem – the biases in need of correction – points to the solution, to the means needed to solve it.* If power blinded

[9] Haslanger, Sally. 1993. "On Being Objective and Being Objectified." In *A Mind of One's Own*, eds. Louise Antony and Charlotte Witt. Boulder, CO: Westview Press, 85–125.

[10] See Ross, Lee. 1977. "The Intuitive Psychologist and His Shortcomings: Distortions in the Attribution Process." In *Advances in Experimental Social Psychology. Vol. 10*, ed. Roger Berkowitz. New York: Academic Press, 173–220 for a classic discussion. The fundamental attribution error has been subject to refinement. One important refinement enhances its applicability to the case at hand: findings that the powerful are more subject to attribution biases than the powerless. See Fiske, Susan. 1993. "Controlling Other People: The Impact of Power on Stereotyping." *American Psychologist* 48: 621–8.

[11] Hammond, James Henry. 1858. "Speech in the Senate, 35th Congress, Session 1." *Congressional Globe*, 4 March, Appendix, 68–71.

[12] I stress that slaveholders and their many allies sincerely believed that slavery was morally right. They were not like the criminal who knows perfectly well that robbery is wrong but robs people anyway out of self-interest. What needs to be explained is not self-interested wrongdoing but the failure to grasp that what one is doing is wrong. Slavery advocates were astonished by the abolitionists' moral condemnation of slavery. See Stringfellow, Thornton. 1860. "The Bible Argument, or Slavery in the Light of Divine Revelation." In *Cotton Is King, and Pro-Slavery Arguments*, ed. E. N. Elliott. Augusta, GA: Pritchard, Abbott & Loomis, 519 for a characteristic statement.

people to the suffering of others, then taking them down a notch could help open their eyes and hearts, and publicizing the testimony of the suffering could draw attention to what they needed to see and feel. If institutionalized power to get what they want insulated people from recognizing their duties to others, then disrupting the ability of those institutions to deliver what they want, in conjunction with practices of moral blame and accountability, could induce the experiences that help people distinguish what they want from what is right. If dominating power enables objectification by constraining the objectified to act only in accordance with degrading accounts of their innate capacities, then liberating the objectified to act otherwise, and enlisting them in projects where they could manifest superior capacities, could force a reassessment of their natures.

Antislavery activists adopted all of these bias-correcting strategies. Abolitionists in Britain and the United States formed an entire repertoire of practices to actively contest slavery, including mass petitions, election campaigns, lobbying, and boycotts of slave-grown sugar.[13] They did not only make moral arguments. They also held slaveholders to account in litigation against slavery and in exposing slaveholders and their supporters to blame for their cruelty and injustice. Two points are of special interest for pragmatist methods. First, abolitionists grasped, perhaps inchoately, that morally objectionable practices need to be opposed in action, not just in arguments. Armchair thinking alone does not correct moral biases. Action in the world is often needed to supply the circumstances prompting moral re-thinking, to generate new evidence for moral reflection, to counteract or undermine biased thinking, and to alter the range of alternatives under serious consideration. Such action challenging status quo arrangements, including the ideologies that rationalize them, is called "contentious politics" and consists in coordinated action by groups around a shared agenda, which often aims to disrupt the routine operation of challenged institutions.[14]

Second, abolitionists also recognized the centrality of participation by slaves and free blacks in political contention over slavery. Slave narratives supplied critical evidence against proslavery writers' claims that blacks were happy under slavery. Some of them, such as Olaudah Equiano's *Interesting Narrative*, the key slave narrative circulated by British abolitionists, devoted more time to Equiano's experiences as a freed person than as a slave, simultaneously testifying to and manifesting the powers proslavery writers insisted that blacks lacked – intelligence, self-command, erudition, manners, willingness to work hard, and desire for freedom.[15] Equiano himself campaigned against slavery, demonstrating his superior capacities. Again, deeds by slaves and freed people, more than words about them, mattered. The very fact that slaves yearned for freedom so

[13] Hochschild, Adam. 2005. *Bury the Chains: Prophets and Rebels in the Fight to Free an Empire's Slaves*. Boston: Houghton Mifflin.

[14] Tilly, Charles, and Sidney Tarrow. 2006. *Contentious Politics*. New York: Oxford University Press.

[15] Equiano, Olaudah. 2003. *The Interesting Narrative and Other Writings*. Ed. Vincent Carretta. New York: Penguin Books.

much that they were willing to risk their lives in escaping from it demonstrated both the absurdity of proslavery claims about its purportedly benign nature and the courage and tenacity of the slaves.

Toward the end of the Civil War, when the Confederacy was running out of soldiers, its defenders debated whether to enlist slaves in the Confederate army. Against this, Howell Cobb, one of the founders of the Confederacy, argued that "If slaves will make good soldiers our whole theory of slavery is wrong."[16] But slaves *did* make good soldiers, as the Union army was demonstrating every day. Over the course of the war, fugitive slaves and free blacks constituted 10% of the Union army and consistently fought with valor. Nothing could be a more effective refutation of the proslavery argument that blacks were fit only for servitude than their effective contention against slavery as soldiers.

We should not suppose, however, that blacks' contention against slavery was important only for the ways in which it supplied factual evidence against proslavery arguments and in favor of abolitionist views. Even more important was its emotional impact. Frederick Douglass clearly understood this point of moral psychology in arguing that "human nature is so constituted that it cannot honor a helpless man, although it can pity him; and even this it cannot do long, if the signs of power do not arise."[17] The German language captures this insight in the word "achtung," which means "respect" but also "attention" and "danger." *Active resistance against oppressors by the oppressed counteracts the oppressors' power biases by reducing their power.* Nothing inspires contempt on the part of the powerful toward subordinates more than their submission to domination. Nothing undermines that contempt more than their active refusal to submit.

While bias reduction helps achieve greater moral clarity, it does not guarantee sound conclusions. From a pragmatist point of view, the ultimate test of any moral view lies in experiments in living. Thus, the ultimate vindication of anti-slavery lay in the relative success of the alternative labor systems of production and social ordering in the former slaveholding territories, demonstrating that the freed people could govern themselves.[18] In the United States, the successor regime in the South hardly counted as fully free and certainly not as non-racist. Nevertheless, it amounted to a major step forward in freedom and social standing for the former slaves. Sharecropping, the dominant post-emancipation labor regime, secured for the freed people far greater family autonomy and control over their labor than slavery. Its success relative to slavery shattered the

[16] Cobb, Howell. 1865. *Letter to J. A. Seddon, Secretary of War of the Confederate States of America.* Macon, GA, Published in *The American Historical Review* 1.1 (Oct. 1895), pp. 97–8.*

[17] Douglass, Frederick. 1855. *My Bondage and My Freedom.* Ed. James M'Cune Smith. New York: Miller, Orton & Mulligan, 247.

[18] I explain how post-emancipation regimes functioned as an experiment in living to test abolitionist arguments and sometimes were explicitly understood as such experiments, in Anderson, Elizabeth. 2014. *The Quest for Free Labor.* Lecture 9. Amherst Lecture in Philosophy. Amherst College, [[Http://www.amherstlecture.org][Http://www.amherstlecture.org/anderson2014/index.html] .

previously advanced arguments for slavery – for example, that it was necessary for civilization and the survival of the republic and that blacks would starve if not forced to work under the lash. Even in the South, landlords came to defend it.[19]

[19] Ransom, Roger, and Richard Sutch. 2001. *One Kind of Freedom: The Economic Consequences of Emancipation*. 2nd ed. New York: Cambridge University Press, 97.

4 Updating the pragmatist research program

This brief case study of pragmatist methods also illustrates a deep pragmatist theme. We should not suppose that the project of improving our normative judgments, and our practices of reasoning about norms, will end. No immediate post-emancipation regime instituted anything close to what we recognize as free labor today. Notwithstanding the monumental achievements of abolitionist movements, they did not come close to overcoming racism or racist thinking in former slaveholding multiracial societies. Nor should we suppose that, just because we embrace the conclusions of the abolitionists, their original arguments for their views were all sound. How many today think that the case against racialized slavery is clinched by the claim that God "made of one blood all nations of men"?[1] Even our normative concepts change over time, as we confront new circumstances and refine our concepts to cope with them.

Moral arguments have important uses in informing deliberation. But thinking only in armchair, a priori mode – in isolation from practice, from scientific knowledge about thinking itself, and from other people, particularly those affected by our conduct – cannot take us far. To improve our practical reasoning, we need to engage all three. Let us consider each in turn.

1. *Practice.* What does it mean to recommend that we "do what works"? It means to adopt a problem-centered mode of thinking, in which the standards of success are defined contextually, in terms of the problem to be solved. The task is to describe the problem at hand with enough specificity that the statement of the solution is included in the statement of the problem. The description is only provisional. Whether it is vindicated depends on the results of putting the postulated solution into practice and seeing whether it solves the problem as originally articulated, with acceptable side effects. For any complex social problem, tracing causal connections between implemented supposed solutions and problems will often require empirical research drawing upon one or more sciences. The progress of the sciences thus goes hand-in-hand with progress in normative thinking, as we advance our ability to subject our practices to test in experiments in living.

2. *Cognitive science.* The more we empirically investigate human thinking, the more we discover its systematic biases. But we also discover the conditions under which

[1] Acts 17:26.

thinking can succeed, or at least mitigate, its biases. Research into what measures can correct biases in thinking is thus indispensable to improving our methods of practical reasoning.

3. *Democracy.* Most practical problems involve multiple people, who may be differentially affected by the problem and also have different knowledge, skills, and perspectives on it. Forging solutions to problems thus requires engaging them in discussion of what the problems are and what the solutions might be. This, roughly speaking, is the practice of democracy. Improving our democratic practices – our practices of reasoning together – is therefore key to improving our methods of practical reasoning.

To improve our practical reasoning, then, we need to get out of our heads (beyond pure a priori reflection) and into the world – to test our normative ideas in practice, engage the empirical sciences, and engage with other people. These are the keys to being a pragmatist.

6. What is it to be a rational agent?

Ruth Chang

What is it to be a rational agent? The orthodox answer to this question, at least among analytic philosophers, can be summarized by a slogan: *Rationality is a matter of recognizing and responding to reasons.*[2] When you wake every morning, for example, you have reasons to have a balanced breakfast, be on time for work, and be respectful to the people you encounter throughout your day. You may also have reasons to believe that the sun is shining, that if you don't hurry you'll be late for work, and that SARS-CoV-2 virus is here to stay. Being a rational agent is primarily a matter of recognizing these reasons and then responding to them in the right way – for example, by having a balanced breakfast and believing that you'll be late if you don't hurry.

This orthodoxy seems both plausible and attractive. After all, what else could the central activities of rational agency be? Of course, the slogan states the view in its barest, skeletal form, and there is more to being a rational agent than simply recognizing and responding to reasons. You have to have certain other capacities, too, like a decent memory and the ability to make inferences.[3] But the job description of being a rational agent consists primarily in these two tasks: *recognition* and appropriate *response*. Indeed, the bulk of philosophical work about rationality and reasons can be seen as trying to clarify and deepen these two main components of rational agency.

Some questions philosophers have tried to answer concerning *recognition* include: How do we come to recognize a reason? What are reasons? What is the nature of the normativity of reasons? Does recognizing a reason require believing that it is a reason or does it involve having some other form of acquaintance with it? How does recognizing a reason get a 'grip' on rational agents – motivationally or by having rational authority? Is the reason that is recognized *evidence* for something? Does recognition require deliberation, if only implicit? How does automatic action or habit, which accounts for roughly 43% of our movements as rational agents, figure in an account of rational agency?[4] Is there a special faculty of intuition by which we come to recognize something as a reason?

And some questions philosophers have tried to answer concerning appropriate *response* include: What is it to respond to a reason? Is it a matter of having certain mental states such as intentions that are caused or otherwise responsive to the reason? Does responding to a reason involve mechanistic or sub-personal processes that move

[2] See, for example, Susan Wolf (1990), Joseph Raz (1986, 1990, 1999, 2000), Thomas Scanlon (1998), Derek Parfit (2011), Jonathan Dancy (2000, 2004, 2018), John Skorupski (2010), among many others. Sometimes the orthodoxy is expressed in terms of values or oughts; being a rational agent is a matter of recognizing values/what you ought to do and then responding appropriately to what you recognize. The reasons at issue here are *normative reasons*, and the kind of rational agency of interest is the substantive rational agency relating agents to normative reasons rather than the 'structural' rationality relating an agent's movements of mind.

[3] For further discussion of other, arguably subsidiary, capacities rational agents require, see, for example, Joseph Raz (2011).

[4] www.npr.org/2019/12/11/787160734/creatures-of-habit-how-habits-shape-who-we-areand-who-we-become

us from one mental state to another or from one mental state, like intention, to action? What is it to respond appropriately to a reason? Must we act *for* a recognized reason for our response to be appropriate? What is it to act for a reason? How do we put our reasons together so that we can respond to them appropriately? What normative relations hold among reasons – outweighing, bracketing, cancelling, excluding, defeating, trumping, and more? Do our reasons always tell us what it is rational to do or believe, or do they sometimes 'run out'? If reasons run out, then what?

It is no exaggeration to say that the orthodoxy about rational agency, understood in varying terms over the years, has enjoyed preeminence for at least the past twenty-four centuries. But is it correct? Is being a rational agent centrally and primarily a matter of recognizing reasons and then responding to them appropriately? Or might there be some other central activity of rational agency that the orthodoxy overlooks?

My aim in this chapter is to explore, in a rough-and-ready way, an alternative view of what it is to be a rational agent. My focus will be exclusively on *practical* rational agency and *practical* normative reasons. (There are interesting – and far more controversial – analogues in theoretical rational agency that I leave aside here.) According to this alternative, the primary activities of a rational agent are not simply recognizing and responding to reasons but also, crucially, *creating* them.

This unorthodox alternative takes as its springboard a fundamental complaint against the orthodoxy: it assumes too *passive* a view of what it is to be a rational agent. In particular, the orthodoxy posits a passive relation between the *agent*, on the one hand, and *which reasons she has*, on the other. The agent plays no direct role in determining *which* are her reasons; reasons are all *given* to her and never *created* by her. We might call this the 'Passivist View' of rational agency according to which all normative reasons are given to agents and never created by them. The alternative, 'Activist View,' maintains that a central part of being a rational agent is *creating* reasons. It is this Activist View that I want to explore here.[5]

But can we *create* reasons? I suggest a framework for thinking about reasons, borrowed from metaphysicians, that makes sense of the idea that we can, quite literally, create reasons. I explain *how it could be possible* for rational agents to create reasons and explore some upshots of created reasons for thinking about rational agency and reasons.

Now a number of questions and objections immediately arise. The most obvious include: i) if we can create reasons, then it seems that we can make it true that maiming, murder, and mayhem are justified, but that is clearly mistaken; ii) creating reasons, if we really do literally create them, is an arbitrary matter, but reasons aren't arbitrary in this way; and iii) so-called created reasons aren't created at all but are just ordinary non-created reasons that come about in various ways, such as by our

[5] This distinction between the 'Passivist' and 'Activist' views of rational agency cuts across many other 'active' vs. 'passive' distinctions in the philosophy of practical reason. In this volume, Sarah Buss discusses ways in which an agent can be both active and passive that cut across the distinction drawn here.

adopting policies for acting or deliberating. I have tried to address these particular objections elsewhere (Chang 2013a, 2013b, 2017), and no doubt more could be said about each of them. For now, I must ask the reader to put these and other worries aside, as my aim here is not to provide a defense of the Activist View. Instead I want to suggest a positive consideration for thinking that the Activist View might be correct.[6] If we assume, by hypothesis, that the Passivist orthodoxy is true, we quickly run into the problem of how rational agents can appropriately form tractable or 'well-formed' choice situations within which they can contemplate alternatives and choose among them. By allowing that we have the power to create reasons, we have an attractive solution to the problem, a solution that, I argue, is more attractive than those that the Passivist orthodoxy can offer.

[6] Other considerations in favor include ways in which created reasons can (i) answer two puzzles about rational choice (Chang 2009), (ii) explain the special reasons we have in committed relationships (2013b), and (iii) provide us with an account of 'hard choices' (Chang 2017).

1 A grounding framework for reasons

Suppose that you are good with numbers. Suppose too that your being good with numbers is a reason for you to pursue a career in accounting. In virtue of what is your being good with numbers a reason to do so? What makes this rather ordinary consideration about you a reason for you to pursue accounting as opposed to, say, mountaineering or interior design?

Normative reasons have *grounds;* there is something in virtue of which a consideration, like being good with numbers, is a reason for you to do something, like pursue accounting. Metaphysicians propose that the *ground* of a fact is some consideration(s) in virtue of which that fact holds.[1] We can suppose that the ground of a fact in some sense *explains* the fact. What grounds the fact that p and q are the fact, p, and the fact, q. Those two facts explain why the fact that p and q holds. What grounds the fact that the substance in the glass is water is the fact that the substance in the glass has the chemical composition of H_2O. Having that chemical composition is that in virtue of which the substance is water and explains why the substance is water.

With respect to the normative, there are two main grounding relations. One holds within first-order normative theorizing, as when we might claim that the ground of being the right action is the fact that the action maximizes happiness for the greatest number. The fact that it maximizes happiness explains why the action is right. Another holds within metanormative theorizing, as when we might claim that what grounds the fact that something is a reason is God's command that it be a reason. God's command that being a hoofed animal is a reason not to eat it explains why being a hoofed animal is a reason not to eat it. Our interest here is in the metanormative relation of grounding and, in particular, in the question, 'In virtue of what is some consideration, such as being good with numbers, a normative reason for action, such as pursuing a career in accounting?' We seek an explanation of why some consideration is a reason.

Grounding theorists allow that multiple considerations may ground a single fact. Any single consideration that is part of a set of considerations that ground a fact is a 'partial ground' of that fact. The fact that p is a partial ground of the fact that p &

[1] See Fine (2001) for the canonical contemporary statement of the notion of grounding; see also Schaffer (2009) and Rosen (2010). For an overview of the notion, including its history, see Raven (2020). I assume that 'considerations', which include facts, are grounds and that what they ground are 'facts'. There are various niceties concerning ground that need not trouble us here.

q. Some partial grounds are 'primary' or critical to the explanation of the fact, while others are 'background' partial grounds, or, to borrow from normative explanation, 'enablers'.[2] The fact that all crows are black is grounded in the fact that each crow is black but also, it might be thought, in the fact these instances of crows make up the totality of crows. While each fact is a partial ground of the fact that all crows are black, the facts that each crow is black are the primary facts in virtue of which all crows are black. The 'totality' fact is what we might think of as a background or 'enabling' condition that makes the primary grounding facts able to ground the fact that all crows are black. With respect to our metanormative grounding question, we are interested in the 'primary' partial ground in virtue of which a consideration is a reason, allowing that there may various other partial grounds that play enabling conditions. To mark our interest, we can call the primary partial ground of a reason its 'source'.

[2] In the context of first-order normative explanation, see Jonathan Dancy (2004: 38–39).

2 The sources of normative reasons

Return to the fact that you are good with numbers. By hypothesis, this fact is a reason for you to pursue accounting. What is the source of this reason – in virtue of what is it a reason? Philosophers have offered three broad answers to the metanormative question.[1]

The *source externalist* maintains that your being good with numbers is a reason to pursue accounting in virtue of something normative, such as the value of an accounting career for someone who is good with numbers. The 'source' of your reason is given by some external normative fact, such as the goodness of pursuing accounting if you're good with numbers. (It may also be a reason to pursue mathematics, applied physics, or engineering, but we'll stick with accounting.) Some source externalists allow that the explanation of why a consideration is a reason bottoms out in the normative fact that that the consideration is a reason. It's just a normative fact that being good with numbers is a reason to pursue accounting, and there is no more explanation to be had. In this case, the reason is ungrounded, or, for our purposes, we can say, loosely, that it is grounded in a normative fact – the normative fact that the reason is a reason. Source externalists include Plato, Aristotle, Samuel Clark, Henry Sidgwick, H. A. Prichard, G. E. Moore, David Ross, Jonathan Dancy, David Enoch, Thomas Nagel, Derek Parfit, Joseph Raz, Thomas Scanlon, Russ Shafer-Landau, Jonathan Skorupski, Judith Jarvis Thomson, Jay Wallace, and Ralph Wedgwood.

The *source internalist*, by contrast, maintains that a consideration is a reason in virtue of some internal mental state, such as a desire, aim, or goal. Suppose you have a desire to spend your waking hours working well with numbers. According to the source internalist, the fact that you are good with numbers is a reason for you to pursue accounting in virtue of your desire to spend your days working well with numbers. Your desire explains why being good with numbers is a reason for you to pursue accounting. More precisely, what explains why your being good with numbers is a reason to pursue accounting is the *relation* between being good with numbers, the action of pursuing a career in accounting, and your desire to spend your days working well with numbers: being good with numbers is a reason to pursue accounting in virtue of the relational fact that doing so under the condition that you are good with

[1] Even without the contemporary notion of ground to hand, a long line of philosophers have sought to *explain* in an asymmetric, not merely modal and noncausal way, *why* something is a reason, an obligation, a value, or what one should do in terms that go beyond first order normative theorizing. Some have thought that the notion of ground goes at least as far back as Plato and Aristotle (Correia & Schnieder 2012).

numbers will satisfy your desire to spend your days working well with numbers. Some leading source internalists include David Hume, W.D. Falk, the early Philippa Foot, Bernard Williams, Peter Railton, Richard Brandt, Steven Darwall, Kate Manne, Julia Markovits, Shaun Nichols, John Rawls, Michael Smith, Mark Schroeder, Michael Smith, David Sobel, Sharon Street, and Valerie Tiberius.

Source internalism is so called because the source of your reasons is internal to you – given by your desires. Similarly, source externalism is so called because the source of your reasons is external to you – given by facts in the normative world. But this contrast between a source being internal as opposed to external to you obscures another important distinction, that between sources over which we have no volitional control and those over which we do. It is this latter contrast that marks the distinction between reasons *given* to us – whether their source is external or internal to us – and reasons we *create*.

Like normative facts, which ground externalist reasons, the fact that we desire something is not something over which we have volitional control. To see this, consider an example adapted from Elizabeth Anscombe. Suppose I offer you a million dollars if you want to eat a saucer of mud *for its own sake*. Try as hard as you might, you are unable to want this. You can of course want to eat it for the sake of the financial bonanza on offer, and you can undertake measures to cause yourself to form the desire – undergo hypnosis, say – but you cannot *as a matter of will* desire to eat the saucer of mud for its own sake. This is because desires are not under our volitional control; they are just things that happen to us, 'assailing us, unbidden'. The unlucky amongst us find themselves with desires to spend beyond their means, to engage in sexual activities with the dead, and even to murder their enemies. The desires we have are a product of our causal paths and native psychology but not something we can will into or out of existence.

Reasons whose sources are not under our volitional control are *given* reasons; they are given to us, not created by us. Their grounds are in normative facts or desires you happen to have, not in your own willing. The Passivist View of rational agency assumes that all normative reasons are given in this sense. The Activist View, by contrast, maintains that some or perhaps all reasons are created: their grounds are under our volitional control. The Activist View rejects source externalist and source internalism as global views about the source of normative reasons. In its most plausible form it is hybrid; it allows both given and created reasons.[2] To see how normative reasons can be created, we must look to a third possible answer to the metanormative grounding question.

[2] I moot such a view in Chang (2009, 2013a, 2013b, 2017), where I favor a hybrid of externalist given reasons and created will-based ones.

3 Source voluntarism about normative reasons

'Source voluntarism' maintains that a fact is a reason in virtue of the agent's volition – her willing something. Your facility with numbers might be a *given* reason to pursue accounting in virtue of the goodness of pursuing accounting under this condition or your desire to spend your days successfully manipulating numbers. But it might *also* be a reason in virtue of your volition. In this case, since the ground of your reason – your willing something – is a matter over which you have volitional control, you have volitional control over whether you have this reason. We can call reasons whose normative source is in willing 'will-based' reasons. By willing this rather than that, you *create* one will-based reason rather than another. We can, then, quite straightforwardly and literally *create* reasons – will-based ones – by willing their grounds. Source voluntarism is the key component of the Activist View of rational agency.

It is worth mentioning two theories that might be thought to be source voluntarist but are not. One is Kantian. According to Kantian and neo-Kantian views, the reasons you have to act are constrained by the supreme principle of morality, the Categorical Imperative. This supreme principle is the law governing the will. And so it might be thought that Kantians ground reasons in the will. But this is not so. There is a sense in which your reasons derive from your will, but crucially, what you will is not under your volitional control in the sense at issue. Your will is bound by the Categorial Imperative; it's not up to you to will this as opposed to that and thereby it's not up to you to create this reason rather than that. Put another way, we have no volitional control over our willing; what we will is determined by the law governing our will, the supreme principle of morality.[1] Kantians offer two explanations of the grounds of this principle. Some, like Elizabeth Anderson (1993) and Barbara Herman (1993), explain the Categorial Imperative in terms of the intrinsic value of persons or of humanity as such. Others, like Christine Korsgaard (2008, 2009), argue that the law governing the will is a constitutive feature of action. Either way, we cannot *create* reasons; our

[1] Some neo-Kantians allow that you can make a contribution to how your will conforms to the Categorial Imperative by, for example, having a practical identity that guides the reasons you have (Korsgaard 1996) or by choosing things that are good for you but not required by the Categorical Imperative (Hill 2002: 260ff). However, no neo-Kantian maintains that the reasons you have are up to you in the way that source voluntarism maintains – as a matter of your own will, ungoverned by some principle or law. Whether Kantian-inspired accounts belong to externalist, internalist or some further sui generis category of explanation of reasons depends on the details of how such accounts are developed.

reasons are given to us by the intrinsic value of persons or by a normative principle, the Categorical Imperative, that is a constitutive governing principle of action. In this (admittedly, not uncontroversial) way of understanding Kantian views, reasons have their source not in something over which we have volitional control but are given to us.

A second kind of theory that might be thought to be voluntarist is existentialism. But existentialists do not think that we create reasons; on the contrary, they think that the very idea of reasons and rational agency is a chimera. Existentialists, on standard interpretations, are thorough-going nihilists about reasons, rationality, and value. Human life is marked not by rational agency but by *existential* agency. To be an existential agent is to understand that there are no normative reasons that exist antecedent to what we choose to do; each of us simply chooses, unguided by reasons. Of course, after we choose, we may project onto the world what appears to be a normative structure of 'reasons' and 'values' in terms of which our choices can be made intelligible to ourselves and others. But what we spread onto the world has no genuine normativity; the 'reasons' and 'values' to which we may appeal to make sense of our choices are not in themselves binding on us. In short, existentialists are not source voluntarists about reasons since they eschew the very idea of a normative reason.

The source voluntarist maintains that the activity of willing itself is that in virtue of which something is a reason. Since our willing is something over which we have volitional control, by willing something we can, quite literally, *create* a reason. If I will A, I create reason B. If I will C, I create a different reason, D. We create reasons by having volitional control over the source of those reasons. The Passivist orthodoxy recognizes only reasons that are given to us, not created by us. The Activist View, by contrast, allows that part of being a rational agent is creating will-based reasons.

4 What do we will when we create reasons?

The Activist View maintains that rational agency involves the power to create reasons by willing something. But what do we will when we create a reason? I suggest that by willing some fact, F, to be a reason, we thereby make F a reason by way of that very willing. More precisely, we create a will-based reason R to φ, by willing that some fact, F, be a reason, R, to φ, by way of that very willing. Since the reflexive nature of the willing is a detail that is not important here, for convenience we can say that we create will-based reasons by willing *that they be reasons*.

A rough analogy will help. Willing something to be a reason is akin to stipulating the meaning of a word. Take the nonsense word 'corisplay'. You might stipulate that 'corisplay' means 'the sound of leaves rustling in the wind'. (The Greek word 'psithurism' has this meaning, but no English word does). By stipulating that in English 'corisplay' means 'the sound of leaves rustling in the wind', you thereby make it the case that 'corisplay' has this meaning (and is thereby a synonym of the Greek word 'psithurism.') By your linguistic stipulation, you confer meaning on this expression, and it thereby has that meaning.

Willing something to be a reason is what we might think of as *normative stipulation*; by willing some consideration to be a reason, you confer on it the normativity of a reason, and it thereby has that normativity. By stipulating that something is a reason, it thereby becomes a reason, since your willing-it-to-be-a-reason is that in virtue of which it is a reason.[1] The activity of normative stipulation – willing a consideration to be a reason – is the source of that consideration's being a reason.

Moreover, like linguistic stipulation, normative stipulation confers normativity *only for you*. When you stipulate the meaning of 'corisplay', that is its meaning only for you; you can't make that nonsense expression have that meaning for me. Just as you can't create meaning for others, you can't create reasons for others. Of course, when you go around using 'corisplay' to refer to the sound of rustling leaves in the wind, your friends might start using the word in that way too. There can be downstream effects of stipulated meanings that end up resulting in non-stipulated meaning. (Perhaps all neologisms are generated in this way.) Similarly, when you normatively stipulate

[1] An alternative way to understand the analogy is to think that by linguistic stipulation, you assign an expression meaning, and that by normative stipulation, you assign normative reasons-giving force to a consideration.

something to be a reason, you create a will-based reason, and having that will-based reason may have downstream normative effects; if, for example, you act on that will-based reason, you may now have given reasons you would not have otherwise had. Compare you and your Doppleganger, both contemplating whether to pursue a career in accounting. You both have the same *given* reasons to pursue accounting, but if you normatively stipulate that your facility with numbers is a reason for you to be an accountant while your Doppleganger does not, you will have more reasons, all things considered, to pursue accounting than your Doppleganger.

Now for a significant difference. When we engage in linguistic stipulation, we can be understood as adopting what Michael Bratman (1987, 1999, 2007, 2018) calls a 'policy'. When you stipulate that 'corisplay' means 'the sound of leaves rustling in the wind', you decide, going forward, to treat 'corisplay' as meaning 'the sound of leaves rustling in the wind'. Although there is a sense in which, without retracting your stipulation, you make a mistake if you take 'corisplay' to mean something else, the mistake need not be one of failing to follow your normative reasons. The normativity at play when you stipulate the meaning of a word is of the same sort as when you stipulate the rules of an invented game – you make a mistake when you violate your own stipulations, but only relative to the game, which you may have no normative reason to play.[2] When we normatively stipulate a reason, by contrast, we quite literally create a normative reason. If we fail to give that reason the normativity it is due, then we are rationally criticizable for failing to respond appropriately to our normative reasons.

The question naturally arises: Why does willing (that something be a reason) confer the normativity of a reason as opposed to the lesser normativity relativized to a game or mere standard? This leads to a key feature of will-based reasons. I suggest that your willing that something be a reason can be that in virtue of which that thing is a reason because willing-something-to-be-a-reason involves *putting your very agency* behind its being a reason. Put another way, willing that something be a reason is *committing* your very self to its being a reason. When you commit to something's being a reason, you stand behind its being a reason. Normative stipulation differs from linguistic stipulation because the commitment involved in normative stipulation is the special commitment of one's very agency: you stipulate with your very self, as it were, that your talent with numbers is a reason for you to pursue a career in accounting. By putting your very self behind your skill as a reason to pursue this particular sort of career, you give yourself a reason to pursue that career.[3] Typically, in cases of linguistic stipulation, there is no

[2] You might make a mistake of what Thomas Scanlon (1998) usefully calls 'structural rationality', the rationality governing proper movements of and collections of attitudes of the mind. Structural rationality is the kind of rationality at stake when we take your attitudes, policies, and plans as given and evaluate whether they properly belong together and support new attitudes.

[3] Again, willing something to be a reason is not a Bratmanian self-governing policy; such policies are not grounds of normative reasons but elements of an explanation of how an agent acts (see Bratman 2007, 2018).

such agential commitment.[4] Thus, according to the unorthodox view, being a rational agent involves putting your very self behind considerations and making them reasons. It understands rational agency as *active* in this deep way*;* rational agents not only recognize and respond to reasons, but they engage their very selves in the activity of creating them.

[4] You could, of course, commit – in the sense of interest – to your linguistic stipulation that 'corisplay' mean 'the sound of rustling leaves' to be a reason to undertake the action of using 'corisplay' in this way.

5 Some upshots for normative reasons and rational agency

The possibility that some consideration, such as your ability with numbers, is both a given reason and a will-based one suggests a revision in orthodox practices of individuating reasons. According to the orthodoxy, reasons are individuated simply by their 'content', that is, by the consideration that is the reason. Once we allow that something can be a reason in virtue of two quite different sources, it makes sense to individuate reasons not simply by their 'content' but also by their source. In this way, the very same fact – that you are good with numbers – can be both a given and a will-based reason.

We could, of course, continue to individuate reasons by their 'content', but in this case, we must allow that a single content could have normativity that derives from two sources. Rational agents can endow a reason, which might already have one source of normativity in a normative fact or desire, with 'additional' will-based normativity. Your talent with numbers is a given reason to pursue accounting. By putting yourself behind your talent being a reason to pursue accounting, you endow that given reason with additional normativity whose source is your commitment. The Activist View, then, can be understood either as allowing for will-based reasons in addition to given ones, *or* as allowing that the normativity of a reason can be in part 'given' and in part 'will-based'.

The existence of will-based reasons also makes plain that reasons need not be reasons for everyone. Like ordinary agent-relative given reasons, will-based reasons are reasons *only* for the agent who creates them. But they are 'agent-relative' reasons in an extended sense: the *source* of the reason essentially implicates the volitional activity of the agent for whom they are a reason.

What about disagreement about reasons? Does the existence of your will-based reason to pursue accounting preclude disagreement about the reasons you have? You and I can 'disagree' about the reasons you have to pursue a career in accounting. But the 'disagreement' has to be understood in the right way. Suppose your given reasons do not give you most reason to pursue a career in accounting. But you create a will-based reason for yourself to do so, and now, all things considered, you have most reason to pursue accounting.[1] I might say to you, 'You're wrong to think you have

[1] Elsewhere (Chang 2009, 2013a, 2013b, 2017, 2020), I argue that your given reasons must 'run out' in order for you to be able to create a will-based reason that then can make it true that, all things

most reason to pursue accounting'. We can understand what I say in four ways. First, I can be understood as saying, truthfully, that you do not have most given reasons to pursue accounting. In this case, there is no genuine disagreement. Second, suppose we have both drunk the KoolAid and are on board with will-based reasons. I might fail to recognize that you have created a will-based reason for yourself, in which case what I say is false. But, third, and more interestingly, suppose I recognize that you have created a will-based reason for yourself to pursue accounting. If I nevertheless say that you're wrong to think that you have most all-things-considered reasons – given and will-based – to pursue a career in accounting, I might be saying something like: 'If I were in your shoes, I would not have created a will-based reason to pursuing accounting – you shouldn't have normatively stipulated such a reason at all or perhaps you should have normatively stipulated that your facility with numbers is a reason to pursue engineering instead.'[2] While recognizing that you have created a will-based reason for yourself to pursue accounting and that therefore you have such a reason, I can nevertheless judge that your creating this reason was misguided or imprudent or silly. I make these judgments about your reasons from my own point of view – how I would see things were I in your shoes. Finally, I could be saying that you were mistaken in creating your will-based reason from some other point of view, for instance, from the point of the view of the universe (if there is such a point of view). Perhaps your becoming of an accountant precludes a counterfactual reality in which you become an engineer instead and help prevent the rise and dominion of AI as our overlords.

Will-based reasons can be *evaluated* by reasons, but their creation cannot be *guided* by reasons. This distinction is of the utmost importance. When you create a reason for yourself to pursue accounting, you are not *guided* by reasons; creating will-based reasons is not itself an activity guided by reasons. In this way, creating reasons is a bit like sneezing; sneezing is not an activity *guided* by reasons. But we can evaluate the reasons for and against your sneezing – your sneezing was a bad thing because it startled the autocrat into pressing the launch button and starting a war. Although your creating a will-based reason to pursue accounting is not itself guided by reasons, I can evaluate your creation – I wouldn't have done that if I were in your shoes. Crucially, however, although I may think it is a mistake for you to create the reason you did, from your point of view, your activity of creating that reason is not itself guided by reasons. Creating reasons is just something you do.

This leads us to the deepest difference between Passivist and Activist Views of rational agency. The Passivist View maintains that the central activities of rational

considered, you have most reason to do what you created a will-based reason to do. The paradigmatically interesting case in which reasons 'run out' is when they are on a par. But even eschewing parity, reasons 'run out' more often than not; so long as the reasons are 'incommensurable', that is, cannot be measured on a cardinal scale of normative force or importance, rational agents have the capacity, metaphysically speaking, to create reasons.

[2] This way of thinking about will-based reasons sidesteps the problem of accounting for disagreement in cases where the evaluation of a claim is sensitive to context. Cf., for example, MacFarlane (2014).

agency involve being *guided* by reasons. As a rational agent, your job is to recognize reasons, and then to respond to them appropriately, where your response is appropriate insofar as it is guided by the reasons you recognize.[3] Being rational is just a matter of following – being guided by – the reasons given to you. The Activist View sees rational agency quite differently. Being rational centrally involves an activity that is *not guided by reasons*. Being rational has at its very core the capacity to *create* reasons for oneself to, for example, pursue one career over another. Being rational involves being able to put your very self behind something's being a reason and thereby making it a reason for yourself. Creating reasons is part of what rational agents do.

The idea that at the very center of rational agency is the activity of creating reasons, an activity that is not itself guided by reasons, may seem strange. But consider rational agency without such a normative power. Rational agency would be a matter of following, a bit like an automaton, the reasons given to you. Sure enough, some of your given reasons will depend on your desires and psychological makeup, but as we have already seen, your desires are not under your direct volitional control. And some of your given reasons will depend on what you do; if you punch someone in the nose, you now have a given reason to make amends. But even here, you have no control over what reasons you have consequent upon your actions. You don't get to decide which reasons you have after you punch someone in the nose; reasons you have following your actions are still scripted for you (Chang 2020). According to the Activist View, we have the normative power to create reasons for ourselves. But the conditions must be propitious. Rehearsing these conditions would take us too far afield (see e.g., Chang 2013a, 2020). Instead, we end by examining one particular problem for the Passivist View to which, I believe, the Activist View can provide a satisfying and attractive answer.

[3] Guidance by reasons arguable holds of automatic, voluntary action, too. A tennis pro, without deliberation, recognizes and responds to reasons to flick her wrist like so in certain circumstances. A reason for thinking that even automatic action is guided by reasons is the phenomenology of failure to be so guided. In a post-mortem of a match, her coach will point out the reasons she had to flick her wrist like so.

6 The problem of well-formed choice situations

Right now, you are lounging on your living room couch, reading this chapter. What reasons do you have, and how should you respond to them? You might think that you have reasons to continue reading, reasons to get up and stretch, reasons to read a novel instead. Or maybe you have reasons to go to the grocery store, pick up your child from soccer practice, fold the socks in your sock drawer, scrub the bathroom floor. To come to think of it, you might also have reasons to get out your checkbook and make a donation to Oxfam or, indeed, get on a plane to volunteer to help in a place where there are many malnourished children. And so on.

How does the Passivist View suggest you proceed? There are two main approaches. On the 'one-tier' approach, the reasons you have and should therefore recognize are all the considerations that count in favor of (or against) *every* action you could possibly perform right now. This way of understanding rational agency arguably lurks behind classic forms of utilitarianism and other crude forms of consequentialism. Bentham (1970: 38ff) thought that the measure of the utility of a possible action depends on the pleasure and pain it would produce not only across space but also from the present into the future. The reasons you have right now are given by the utility of each of the possible actions you could perform right now, and the one that maximizes happiness for the greatest number is what you have most reason to do. Since the utility of saving lives is almost always greater than the utility of reading philosophy, the appropriate response to your reasons right now would be to put down this paper and start saving some lives. One-tier thinking about rational agency casts a long shadow across contemporary arguments about how we should conduct our lives. It underwrites, I suspect, the influential arguments of Peter Singer, according to which those who are well-off should give away a large portion of their wealth to the needy, and recent corollary arguments of effective altruists, according to which we should direct our resources in ways that help other people the most.

There is much to say about the one-tier approach. But for present purposes, we can set it aside in light of a prima facie difficulty. Human rational agents don't have the capacity to identify all the actions that could be performed at a given time, let alone to consider all the reasons for and against each possible actions.[1] An approach to

[1] Bentham himself recognized that it would be difficult for people to follow the dicates of utilitarianism. He nevertheless insisted that considering all possible actions and maximizing utility in our ac-

rationality that makes us irrational all the time is no approach at all. Indeed, intuitively speaking, when we 'choose' between options, we do so among a restricted set of options, against a restrictd background of circumstances, governed by a restricted set of aims, purposes, and values that matter in the choice.

Any attempt to limit the range of actions from which we are to choose involves taking a 'two-tier' (or multi-tier) approach to rational agency. A rational agent recognizes and responds to reasons within a restricted set of circumstances, values, and options, that is, within what we might call a 'well-formed choice situation'. Well-formed choice situations provide a second, restricted tier within which choice takes place. A well-formed choice situation is a circumscribed set of circumstances – as opposed to every circumstance in the world – a relatively small, finite set of alternatives – as opposed to every action you could possibly perform right now, however finely individuated – and a reasonably well-defined set of normative criteria that provide *what matters* in the choice between the alternatives – as opposed to all the values there are. To choose between a relatively narrow set of options, you must be in a well-formed choice situation. If the choice situation is not well formed, your options are not well formed, and you won't be in a position to determine what matters in the choice to begin with. Nor will you be able to justify your choice to continue reading this paper; your choice to do so cannot be justified against a background of severe human suffering that it would be easy for you to help alleviate without your being justified in being in a choice situation in which helping ameliorate suffering is not one of the options.

But now we have a problem. How do you 'get into' one well-formed choice situation as opposed to another?[2] And, a corollary problem, what is your justification for being in one situation as opposed to another?

There are two kinds of answer to these questions, one non-normative and the other normative. Perhaps the non-normative facts – including past actions and human psychology, dispositions, capacities, and limitations – make salient certain choice situations over others. You took the action of enrolling at university, became a philosophy major, and now your philosophy professor has assigned you this chapter as homework. Choice situations in which what matters is completing your homework are especially salient to you. But salience need not determine how you come to be in one choice situation rather than another. You have the capacity to step back and ask yourself, 'Is this the choice situation I should be in?' Moreover, nonnormative facts, even if they fully determine your being in one choice situation rather than another fail to justify your being in that choice situation. Again, you can ask yourself, 'Should I be in this choice situation?' In short, being in a choice situation can itself be a choice.

---

tions was an ideal towards which we all should strive (Bentham 1970: 40). The cursory way in which I treat the one-tier approach should not be taken to signal that I think it can be discarded so easily. I believe it is importantly linked to a common interpretation of the idea that what we should do as rational agents is what we should do *all things considered.* Here, my focus is on the Passivist's best explanation of a 'two-tier' approach, discussed below.

[2] More precisely, there may be range of well-formed choice situations in which one makes a choice,

Perhaps normative facts – like the fact that you have a reason to be in one choice situation rather than another – can determine which choice situation you should be in, and then, as a rational agent, you can respond to those reasons and get yourself to be in that choice situation. Of course, it is not enough that you have a reason to be in a choice situation, to solve the problem of justifying being in a choice situation, you have to have *most* or perhaps *sufficient reason* to be in that choice situation.[3] But how could it be true that you have most or sufficient reason to be in a choice situation in which what matters is finishing your homework when there are countless people suffering in ways that you could help alleviate?

We can begin to answer this question on behalf of the Passivist orthodoxy by affirming that the question, 'Which choice situation should I be in?' presents the agent with a distinctive choice that is governed by *agential values*, such as autonomy, well-being, and meaning in life, where perhaps – we should leave the matter open – these values reduce to facts about what the agent wants. Moreover, this choice of which choice situation to be in need not, of course, be the kind of deliberative and deliberate choice in which one weighs up pros and cons. It is a choice in the broad sense of a being an intentional human action that is guided by and evaluable by reasons.

Now, with respect to these agential values, it becomes unclear whether the appropriate choice situation to be in is always one that prioritizes relieving the suffering of others. Respecting agential values does not require us to maximize utility; such values give us the freedom to choose choice situations in which we can live life autonomously, well, and with meaning. Doing the maximal good for others may not always be the

---

but the differences between situations within this range will not be important for our purposes.

[3] Views in the neighborhood can be co-opted to support the claim that reasons determine which choice situation you should be in. For example, in discussing the difference between 'optional' and 'sufficient' reasons, Joseph Raz (1999:97) writes: "That the chair is comfortable is something good about the chair, and we can say that is a reason to sit on it, but such a reason is not a sufficient reason. If one has reason to rest one's legs then one has a sufficient reason to sit on this chair because it is comfortable." Raz's idea is that some considerations that make action eligible are 'optional' – they aren't 'sufficient' reasons to do something – and can be made sufficient only by another reason. This is not the same idea as but close to the idea that something can be a reason for an action only if it is a reason in a well-formed choice situation in which one has a reason to be. The comfortableness of the chair is a reason to sit in it only if you have a reason to be in a choice situation in which being comfortable matters and sitting in a chair is one of your options. Thomas Scanlon (2014: 106–108) makes similar points about the optionality of reasons. But he goes on to make further illuminating points about the 'weights' of reasons, which can also be coopted to support the claim here. Scanlon (2014: 114) suggests that in ascertaining the 'weights' of reasons, policies and requirements of valuable relationships can play a role. He writes: "In order to lose weight by dieting, or to become healthier through exercise, one needs to have a general policy of giving greater weight to following one's diet or exercise plan than to (at least most of) the considerations of pleasure or convenience that provide reasons for deviating from this plan on a given day." (Scanlon 2014: 114). Moreover, requirements of personal relationships can determine the weightiness of certain reasons: friendship "involves ... taking a certain view of the reasons one has. For example, one would not be a good friend if one did not give priority to one's friend's needs [over personal cost]" (Scanlon 2014: 114, see also Raz 1986: 345–366). We might adapt Scanlon's remarks here to support the idea that policies and requirements of valuable relationships can give us reasons to be in some choice situations over others.

best kind of choice situation to be in with respect to these values, all the more so if these values reduce to facts about what we want. We can suppose that these values make a *range* of choice situations rationally *eligible* to an agent at a point in time.[4]

What could the Passivist say next? If there is a range of eligible choice situations open to an agent at any given time, then the natural thing for the Passivist to say is that the agent has *sufficient* reason to be in any of them. By hypothesis, you have no more reason to be in one rather than another of a range of eligible choice situations. This leaves us with the following result: If someone asks you why you chose to be in a choice situation in which your options were to read some philosophy or get a cup of coffee as opposed to being in a choice situation in which your options were to write a check to Oxfam or get a plane to volunteer your services in an area devastated by a natural disaster, all you say by way of reply is, 'Well, I had sufficient reason to be in either choice situation, and I *just chose* to be in the one rather than the other. And I could have chosen differently, but I didn't'.

This answer need not be troubling for a single choice at a moment in time. But iterated and aggregated over a human lifetime – and across all human lifetimes – we are left with a deeply unsatisfying view of human rationality. You have no more reason to lead your life as a philosopher rather than an architect or lawyer. You have no more reason to spend your life with the man you love, Sam, rather than some stranger, Tom, or to have had children rather than remain child-free. You have no more reason to have devoted your life to working for racial justice than to instagramming the latest runway fashions. At any point in time in your life, many different choice situations are open to you, each of which you have sufficient reason to be in relative to agential values. If there is no more reason for you to be in one rather than another, the trajectory of your life, made up of the choices you make in arbitrarily selected choice situations, is itself profoundly arbitrary. Indeed, we could imagine an AI machine whose job it is to randomly select for every human agent one among the many eligible choice situations open to that agent at each point in her life. Having such a machine determine the choice situations you face throughout your life would be compatible with the Passivist View of rational agency. Rational life would be one very large toss-up.

But this makes a travesty of the human condition. The things that you hold most dear, the things you care about and that give your life meaning, are not things that you have no more reason to have in your life than other things. You would not be content to shrug your shoulders and say, 'Well, I could have married someone else and had different children, but as it happens, I married the light of my life, Sam'. Most people in genuine love relationships correctly think that their relationship is more

[4] Cf Raz's (1997) 'classical conception' of rational agency according to which reasons are considerations that make actions eligible. Here, the idea is that rational agency involves reasons that make choice situations, not actions, eligible. We might go further and say that the choice situations are *on a par* with one another (Chang 2002). Note, too, that, *pace* Raz 1986, eligibility is not plausibly a matter of each choice situation be equally as good as any other with respect to agential values. What I have elsewhere called 'The Small Improvement Argument' makes that clear.

significant or valuable than other relationships they could have had. The value we place on things that enrich our lives and give it meaning is in tension with the view that we have sufficient reasons to be in any of a wide range of sets of choice situations, none of which we have more reason to be in rather than any other. The problem with the Passivist orthodoxy is that it gives us no direct control over which choice situations we are in and thus no direct control over the reasons we have in leading our lives.

The Activist View offers an attractive alternative. Rational agents have the normative power to create will-based reasons to be in one choice situation rather than another. By creating a reason to be in one among many eligible choice situations, you create the justification for being in that choice situation rather than the others. And as a rational agent who responds to reasons, you can thereby get yourself into that choice situation since you have most reason to be in it. And, as we've suggested, when you create a reason for yourself to be in one choice situation among others, you put yourself behind that reason. By putting yourself behind that reason, you make yourself into the kind of person who now has most reason to be in that choice situation rather than any others. In this way, the activity of your will allows you to *become* one kind of agent rather than another, namely an agent who faces *these* choice situations and not *those*. You are the driver of which choice situations – and consequently which reasons – make up the story of your life.[5] By creating reasons for yourself, you form what I have elsewhere called your 'rational identity' (Chang 2009, 2013a).

[5] An alternative, 'Passivist' view of 'becoming' is provided by Aristotelian specificationism (Cf. Richardson 1994, Millgram 2001, and especially Callard 2018): we adopt inchoate, poorly specified values or ends, and part of our agency involves specifying this end and acting on reasons that are implicated in its specification.

7 The activist view in action

Return to you lounging on your living room couch. There are a range of eligible choice situations you could be in right now. This range is determined by agential values like autonomy, well-being, and meaning in life. In choice situation A, what matters is getting your homework done well, and your choice is between continuing to read or getting yourself a coffee. In choice situation B, what matters is the suffering of others, and your choice is between writing a check to Oxfam or hopping a plane to volunteer your aid. In choice situation C, what matters is having fun, and your choice is between going to a movie or calling up some friends for a party. All three choice situations are eligible to you right now.

Which choice situation should you be in? The Passivist orthodoxy has only this to say: you have sufficient reasons to be in any of the three, so *just choose* (or worse, be caused without justification to be in one rather than the other). By hypothesis, there is no reason to be in one over the others. But the reasons that render the choice situations eligible on the Passivist View are *given* reasons. As far as your given reasons are concerned, there is no further justification to be had for being in one choice situation over any others. The Activist View, by contrast, allows that you might create a will-based reason to be in situation A, which then justifies your being in that choice situation. By creating a will-based reason to be in situation A, you thereby make yourself into the sort of person for whom it is true that he has most reason to be in situation A. Your friend, similarly situated, might create a will-based reason for herself to be in situation C. She thereby makes it true of herself that she has most reason to be in situation C. Iterated across a lifetime, you may create a rational identity for yourself as a nerd and your friend a party animal. The Activist View gives rational agents the power to craft their own identities as individuals who justifiably face certain sets of choice situations rather than others.

The path we cut through life, among the myriad choice situations rationally open to us, is justified by the will-based reasons we create. Those who champion effective altruism have cut one such path. Those who spend their hours on Wall Street, making as much money as they can in order to live the high life, have cut another. It is only by allowing that there is more to rational agency than recognizing and responding to reasons that we can make sense of how we can be justified in crafting ourselves into the distinctive rational agents we are. Central to being a rational agent is creating reasons for ourselves to be in one choice situation rather than another. By doing so, we can determine for ourselves the reasons we have.

References

Anderson, Elizabeth. 1993. *Value in Ethics and Economics*. Cambridge, MA: Harvard University Press.

Bentham, Jeremy. 1970. *An Introduction to the Principles of Morals and Legislation*, ed. J. H. Burns and H. L. A. Hart, as part of *The Collected Works of Jeremy Bentham*, eds. J. H. Burns, J. R. Dinwiddy, F. Rosen, and T. P. Schofield. London: Athlone Press; Oxford: Clarendon Press.

Bratman, Michael. 1987. *Intention, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press.

———. 1999. *Faces of Intention*. Cambridge: Cambridge University Press.

———. 2007. *Structures of Agency: Essays*. Oxford: Oxford University Press.

———. *2018.* Planning, Time, and Self-Governance: Essays in Practical Rationality. *Oxford: Oxford University Press.*

Callard, Agnes. 2018. *Aspiration: The Agency of Becoming*. Oxford: Oxford University Press.

Chang, Ruth. 2002. 'The Possibility of Parity'. *Ethics* 112: 659–688.

———. 2009. 'Voluntarist Reasons and the Sources of Normativity'. In *Reasons for Action*, eds. D. Sobel and S. Wall. New York: Cambridge University Press, pp. 243–271.

———. 2013a. 'Grounding Practical Normativity: Going Hybrid'. *Philosophical Studies* 164 (1): 163–187.

———. 2013b. 'Commitments, Reasons, and the Will'. In *Oxford Studies in Metaethics*, ed. R. Shafer-Landau, vol. 8. Oxford: Oxford University Press, pp. 74–113.

———. *2017. 'Hard Choices'.* American Philosophical Association Journal of Philosophy *92: 586–620.*

———. 2020. 'Do We Have Normative Powers?' *Aristotelian Society Supplementary Volume* 94: 280–300.

Correia, Fabrice, and Benjamin Schnieder, eds. 2012. *Metaphysical Grounding: Understanding the Structure of Reality, Introduction*. Cambridge: Cambridge University Press.

Dancy, Jonathan. 2000. *Practical Reality*. Oxford: Oxford University Press.

———. 2004. *Ethics Without Principles*. Oxford: Clarendon Press.

———. *2018.* Practical Shape. A Theory of Practical Reasoning. *Oxford: Oxford University Press.*

Fine, Kit. 2001. 'The Question of Realism'. *Philosophers Imprint* 1: 1–30.
Herman, Barbara. 1993. *The Practice of Moral Judgment*. Cambridge, MA: Harvard University Press.
Hill, Thomas. 2002. *Human Welfare and Moral Worth*. Oxford: Clarendon Press.
Korsgaard, Christine. 1996. *The Sources of Normativity*. Cambridge: Cambridge University Press.
———. 2008. *The Constitution of Agency*. Oxford: Oxford University Press.
———. *2009.* Self-Constitution: Agency, Identity, and Integrity*. Oxford: Oxford University Press.*
MacFarlane, John. 2014. Assessment Sensitivity: Relative Truth and Its Applications*. Oxford: Oxford University Press.*
Millgram, Elijah. 2001. *Varieties of Practical Reasoning*. Cambridge, MA: MIT Press.
Parfit, Derek. 2011. *On What Matters*, vols. I and II. Oxford: Oxford University Press.
Raven, Michael, ed. 2020. *The Routledge Handbook of Metaphysical Grounding*. New York: Routledge.
Raz, Joseph. 1986. *The Morality of Freedom*. Oxford: Clarendon Press.
———. 1990. *Practical Reason and Norms*. Princeton: Princeton University Press.
———. 1997. 'Incommensurability and Agency'. In *Incommensurability, Incomparability, and Practical Reason*, ed. R. Chang. Cambridge, MA: Harvard University Press.
———. 1999. *Engaging Reason*. Oxford: Oxford University Press.
———. 2011. *From Normativity to Responsibility*. Oxford: Clarendon Press.
Richardson, Henry. 1994. *Practical Reasoning About Final Ends*. Cambridge: Cambridge University Press.
Rosen, Gideon. 2010. 'Metaphysical Dependence: Grounding and Reducation'. In *Modality: Metaphysics, Logic, and Epistemology*, eds. B. Hale and A. Hoffman. New York: Oxford University Press.
Scanlon, Thomas. 1998. *What We Owe to Each Other*. Cambridge, MA: Belknap Press.
———. 2004. 'Reasons: A Puzzling Duality?' In *Reason and Value: Themes from the Moral Philosophy of Joseph Raz*, eds. Jay Wallace, Philip Pettit, Samuel Scheffler, and Michael Smith. Oxford: Oxford University Press.
———. 2014. *Being Realistic About Reasons*. Oxford: Oxford University Press.
Schaffer, J. 2009. 'On What Grounds What'. In *Metametaphysics: New Essays on the Foundations of Ontology*, eds. D. J. Chalmers, D. Manley, and R. Wasserman. Oxford: Oxford University Press.
Skorupski, John. 2010. *The Domain of Reasons*. Oxford: Oxford University Press.
Wolf, Susan. 1990. *Freedom Within Reason*. New York: Oxford University Press.

Part 2: Practical reason in the history of philosophy

7. Practical reasoning in early Chinese philosophy

David B. Wong

This essay on practical reasoning in early Chinese philosophy will focus on Confucian and Daoist thinkers. The practical reasoning discussed and displayed in their work has several major features: 1) construing what's right as dependent on situational features in a way that does not rely on general principles; 2) anchored in concrete reflection on particular persons and their actions as exemplars of moral excellence; 3) analogical in the sense of drawing a conclusion about a particular case based on its similarities to other cases in which it is thought that a good conclusion had been drawn; 4) appreciative of the way that reflection and feeling can interweave and mutually influence each other in illuminating ways; 5) having some degree of confidence in intuitive insight, the successes of which cannot be fully explained yet cultivated; and 6) tolerant of apparent contradiction in a way that expresses skepticism about human powers to satisfactorily identify what is good and bad or right and wrong. The second, third and fourth characteristics are most distinctively associated with Confucian practical reasoning, and the sixth with the Daoist approach. The first and fifth are prominent in both.

This chapter explains how these features appear in early Chinese texts and their relationship to each other.[1] Taken together, the six features present an alternative to formal approaches to practical reasoning, an alternative that has arguably been the way human beings most often have navigated through the world, both natural and social. Those partial to the Confucian approach tend to neglect the sixth feature, but incorporating a measured dose of it can help counteract some of the rigidities of perspective to which Confucianism is liable. I will begin with the features of practical reasoning that are most developed in Confucianism and then move on to Daoism.

[1] The reference to "early" Chinese philosophy is to the "pre-Qin" period of Chinese philosophy, 6th century BCE to 221 BCE. This is the era in which ☒☒☒☒ *Baijia zhengming*, or "The Hundred Schools of Thought Contended." A profusion of teachings bloomed to address the great political and social turmoil that occurred during this period, which is contemporaneous with classical Greek philosophy: the pre-Socratics, Socrates, Plato and Aristotle. Because of space limitations, I will not address some of the early Chinese schools that advanced notable theories of practical reasoning. The "later Mohists" (those who developed and extended the primarily ethical thought of ☒☒ or Mozi [Master Mo]) developed a sophisticated philosophy of language centering on ☒ *ming*, or "names," which they held to stand for kinds of things that were grouped together by resemblances. In the ☒☒ *Mojing* or *Mohist Canons*, they held that knowledge derives from the ability to reliably ☒ *bian* (distinguish, through debate) things according to kinds based on similarities and differences. See Graham 2003 under "Further Reading." Another school that deserves attention, related to the later Mohism, is the ☒☒ *ming jia* or "School of Names," which focused on the relation between language and the world. Two of its most prominent thinkers were ☒☒ Hui Shi, the friend and debating partner to Zhuangzi featured in the story about the huge gourds, and ☒☒☒ Gongsun Long, who generated paradoxes that bear resemblance to those of Greek philosophers such as Zeno. See Makeham 1994 under "Further Reading."

Contextual reasoning in Confucianism

In *Analects* (☒☒ *Lunyu*) 4.10, the Master said, "Anywhere in the world, the *junzi* insists on nothing and refuses nothing.[1] He simply goes with what is right." Here going with what is right is associated with the *junzi* ☒☒ or morally noble person, the most prominent character ideal in the text. The word "*yi* ☒," translated as "right," has the connotation of "appropriateness." To go with what is right is not to invariably insist on any general course of action or to invariably refuse any general course. It is to act appropriately in light of the situation or context. "*Yi*" is also applied to a person as a virtue, a reliable sense for what is appropriate and a dedication to acting on what is appropriate. As a virtue, it is typically rendered as "righteousness." 11.22 illustrates the notion of acting appropriately according to context. Zilu asks, "May one immediately practice what one has learned?" Confucius (☒☒ Kongzi) says, "Consult father and elder brothers first." Ranyou asks the same question. Confucius urges him forward. The differing contexts for the differing actions Confucius advises are the weaknesses that each man must address. Ranyou is too cautious and hesitant to express or act on his own views, especially if he fears the displeasure of powerful superiors, while Zilu is too eager to act without exercising enough prior thought or having acquired learning to make his judgments informed. Confucius' different advice to the two men is appropriate in light of their different weaknesses. In the terminology of contemporary ethical theory, Confucius adheres to a moderate version of moral particularism: that moral reasoning is not a "top-down" application of moral principles to particular cases but rather a reasoning from the particular (see McNaughton 1988; Dancy 1993). It is a moderate version because there are generalizations in Confucian reasoning that identify considerations that weigh in moral reasoning, but they must be weighed along with others in a particular context.

How does one develop the ability to determine what is *yi* in a given situation, especially if it is determined according to context? In the *Analects*, the relationship between Confucius and his students illustrates one pathway for the development of this ability: observing models or exemplars of virtue and what they do. The exemplar might be a historical figure, perhaps long dead, but about whom stories are told precisely because of the impact they made in their time and after. The Duke of Zhou, born centuries before Confucius' time, is a founding figure of the Zhou dynasty. He and

[1] All references for the *Analects* are to *Analects* (2017).

his brother, King Wu, worked together to overthrow the corrupt and tyrannical Shang dynasty, but Wu died and his son, the rightful heir, was too young to succeed him. The Duke did not display any ambition to take power for himself but ably administered the country on the young heir's behalf until he came of age. Reflecting on stories told about the Duke might lead one to generalize that those who put their countries ahead of personal gain demonstrate a particularly admirable form of moral excellence. More generally, Amy Olberding (2008, 2013) has argued that the *Analects* displays an "exemplarist" epistemology that starts with particular people and their actions as the epistemic base from which tentative and provisional generalizations about the nature of moral excellence are made.

Reasoning from concrete experience

The central quality of the *junzi*, ☒*ren* (4.5), is presented in the text as something that Confucius and his students are seeking to understand through examining particular people and their actions as possible exemplars of the quality rather than as something that the tradition has already characterized in set terms. The term *ren* has received a variety of translations, for example, "humaneness, humanity, goodness, benevolence." The variety indicates that the *ren* is associated with a wide range of admirable qualities such as loving persons (12.22), self-discipline through observing ritual propriety (12.1), deference, trustworthiness and tolerance (17.6). For this reason, one of the central meanings of *ren* in the *Analects* is that of comprehensive moral excellence, the total package of particular virtues. The varied and unsystematic nature of what is said about *ren* in the text also suggests that the *Analects* is not a systematic presentation of a set view of what *ren* is but rather an exploration of what it is through examination of exemplars. Generalizations have credibility to the extent that they identify something that is pre-theoretically attractive about the exemplars.

Confucius serves as an exemplar to his students, as illustrated by 9.3, where he is asked to decide between contemporary and traditional variants of ritual practices (☒ *li*). In one case, Confucius accepts the contemporary practice because it saves money by using less expensive material for ceremonial caps; in the other case, he accepts the traditional form because it conveys the appropriate attitude of deference to a superior, and the contemporary form conveys arrogance. One might infer, then, that the appropriateness of the attitude expressed by a ritual form has a lot to do with whether the form is appropriate. There are many close observations of how Confucius went about his daily life, and these observations portray a person who is able to express, often wordlessly, his concern and sensitivity to the situation of others (e.g., 10.22). Moral exemplars can also inform others of the larger vision of life that inform their judgments as to what is *yi*. In 5.26, Confucius suggested to two of his students that they each speak of his heart's desire. When they ask Confucius to speak of his, he says that he would like to bring contentment to the aged, to share relationships of trust and confidence with his friends and to love and protect the young. Even if there are no exceptionless rules to guide choice, one can make them while informed by a vision of the kind of life one wants to lead. If it is characteristic of the *Analects* that there is more epistemic confidence in the concrete that is experienced initially at a

pre-theoretical level, the method of analogical reasoning, which is most fully developed in the *Mencius*, makes a lot of sense as a way of extending knowledge from that base.

Analogical reasoning in the *Mencius*

Analogical reasoning starts from a particular case, say, the Duke of Zhou's serving as regent and that it was *yi*. The exemplarist model places the most confidence in such particular judgments. How might one use them as baseline cases to reason about problematic cases in which one is uncertain about what is to be done? The method of analogy starts with identifying cases within one's stock of baseline cases that are the most similar to the present situation. If there are relevant similarities and no relevant dissimilarities between a baseline case A and the present case B, one can seek to identify an action that can be performed in B that is relevantly similar to the action taken in A. It is important to note that identification of relevant similarities and dissimilarities does not depend on having a general principle that identifies them. All it takes is to be convinced that the particular and relevant features of cases A and B are sufficiently similar. In fact, whatever moral generalizations one eventually formulates are derived from repeated instances of analogical reasoning based on similarities between particular cases. Similar inferences can be made that involve baseline cases in which it is clear that the wrong or inappropriate thing has been done.

Some of the clearest cases of analogical reasoning in the *Mencius* (⌧⌧ *Mengzi*) support negative conclusions of this kind, which can in turn support positive conclusions as to what is to be done. In 1A3,[1] King Hui of the state of Liang complains to Mencius that he takes certain actions to alleviate the suffering of his people when the harvest is bad, while kings of neighboring states fail to take similar actions. Yet people in these states do not cross the border to Liang (people were the most valuable resource in the predominantly agricultural economy of China, so kings of the various states were interested in attracting more people from neighboring states). Mencius explains why this has happened through an analogy to an event in war when soldiers turn and run from battle. Some stop running after one hundred paces, and others stop running after fifty. Would it be appropriate, Mencius asks, for the latter to ridicule and criticize the former? In Liang, observes Mencius, the King's pigs and dogs are eating food intended for people while people are dying from starvation on the roads. In addition to giving people the food, King Hui could employ "sustainable resource" measures to help his people feed themselves, such as not allowing closed mesh nets to be used in fishing of

[1] All references for Mencius are to *Mengzi* (2017).

the ponds and lakes and allowing axes to go into the forests and mountains only in the appropriate seasons.

For Mencius, the knowledge embodied in the baseline cases comes from *Tian*, which literally means "sky," but is for him the source of order in the world, including the normative order for human beings. *Tian* implants in human nature cognitive and affective dispositions that can develop into fully realized dimensions of moral goodness or virtues. ☒☒☒☒ *ceyin zhi xin* (compassion) can develop into ☒ *ren* (human-heartedness or the virtue of loving or caring for others); ☒☒ *xiuwu* (shame and dislike) can develop into ☒ *yi* (as an excellence of persons, usually translated into English as "righteousness," an ethical excellence of persons especially featuring a person's dedication to do what is *yi* in the sense of the right or appropriate thing to do); ☒☒ *cirang* (deference and yielding) can develop into ☒ *li* (observing ritual propriety, or observing ceremonies and customary practices that signify respect for others); and ☒☒ *shifei* (this/not this, approval/disapproval, right/wrong) can develop into ☒ *zhi* or wisdom.

The inborn tendency to have compassion for the suffering of others displays itself in spontaneous and unpremeditated responses to the perceived suffering of others, such as a child about to fall into a well (2A6). There is a dimension of Mencian compassion that could be called cognitive: recognition of another as suffering or about to suffer. This dimension is most plausibly construed as constitutive and not merely a causal factor. To regard it as the latter is to presuppose a cleavage of the cognitive from the affective that is just not there in Mencius (see Wong 1991, 2002, 2015). Moreover, it is difficult to distinguish compassion from other feelings in favor of another's welfare without including in its composition a recognition of another's suffering. One of the phrases that Mencius uses to identify the sprout of *ren* is simply "☒☒☒☒ *bu ren ren zhi xin*" or "the heart that cannot bear the suffering of others" (2A6). The reference to alarm and distress also points to a subjectively felt phenomenological element of compassion. Besides the cognitive and phenomenological dimensions, not being able to bear another's suffering involves an impulse to action – an impulse to prevent or alleviate suffering. In fact, what is felt may be the body getting ready for action of some sort. The inborn tendency to perceive and attach importance to what is *yi* displays itself in aversive reactions (including feeling shame or perceiving another's act as shameful) to acts contrary to *yi* such as a hungry person rejecting food that is offered in an abusive way (6A10).

The responses of deference and yielding are the incipient beginnings of the tendency to observe rituals expressing respect for others. The exact relation between the attitudes of deference and yielding, on the one hand, and the excellence of properly observing rituals, on the other hand, is open to interpretation. There is no need to fix a single interpretation to the text. One might hold that the attitudes are specifiable independently of the conventionalized ritualized means of expressing them. Nevertheless, one might hold that expressing the attitudes through such means is a necessary form of ethical excellence because people need common understandings of how they are expressed in a variety of social contexts. That is, they need a common language,

partly given through ritual and other customary practices, to adequately convey these attitudes toward one another. Or one might hold that the attitudes themselves are partly constituted by specifications of what it is to *act* deferentially or respectfully, that these attitudes are not fully specifiable without identifying patterns of conduct one is disposed to display if one had these attitudes.

Mencius does not engage in much explicit discussion of the *shifei* responses, but a reasonable inference from the various ways that *shi* and *fei* are used in the text is that their use is based on inborn dispositions to perceive or judge patterns of similarities and differences that play crucial roles in the cognitive aspects of the other *duan*: who is suffering for example, and which acts are shameful and which are not. When these dispositions develop into abilities to make reliable discriminations of the relevant sort, they help constitute *zhi* as wisdom. In addition, *zhi* seems to involve the more general and sophisticated judgments about the structure of ethical excellence itself. In 4A27, the most authentic expression of wisdom is identified as knowing that the most authentic expression of benevolence is serving one's parents and of righteousness is following one's elder brothers. In 7A46, Mengzi says that what is urgent for the wise is confronting what is fundamental. There is no one whom *ren* persons do not love, but what is fundamental is earnestly caring for the worthy. "Most fundamental" refers to the fact that it is to be given priority in the order of development. One begins with concentrating most of one's efforts to what is most fundamental and extends them to what is less fundamental.

While one might not believe that *Tian* implants inborn beginnings of goodness, there is evidence that human beings might have been endowed during evolution with dispositions that underlie the perception of similarities and dissimilarities crucial to human cooperative life. Consider Wynn and Bloom's studies of young infants who appear able to distinguish "helpers" from "hinderers" and moreover to favor the former and disfavor the latter (Hamlin et al 2007, 2010; Hamlin and Wynn 2011). Or consider Tomasello's work on the readiness of young children (as young as fourteen months) to perceive others as engaged in a task, as needing help, and to jump right in and to help without being asked (Warneken and Tomasello 2007).

Mencian analogical reasoning presents an attractive alternative to those who have become disillusioned with "top-down" models of practical reasoning that deploy extremely general principles (such as maximizing utility or treating each person as an end) to a particular situation in order to judge what is to be done. The analogical approach keeps us closer to the ground level of more specific moral judgments and perceptions, some of which we might have more confidence in than any of the principles of modern moral theory. If one can address a problematic situation by constructing an analogy to relevantly similar past cases in which we have some confidence that the right thing was done, we might arrive at a solution for the situation without having to specify the meaning of a very general principle that covers a vast range of diverse cases. A moderate form of the "top-down" model of reasoning might allow for the use of analogical reasoning when we have confidence in making inferences according to similarities

between cases. Keep in mind, moreover, that using the analogical approach need not exclude deploying general principles when we do have confidence in them or at least in their particular application to the case at hand. Deploying the analogical approach to a related series of cases may even lead us to see that the relevant similarities across these cases might be usefully formulated as principle that identifies types of situational features that weigh in favor of performing a certain type of action. The advantage of formulating principles from an analogical ground is that the level of generality can be adjusted up or down depending on how much confidence we have in the reliability of our identifications of relevant situational feature types. Moderate forms of either model do not exclude one another and might be usefully combined. Alternatively, top-down models that are intended only as theories of right-making characteristics and not as decision procedures would not conflict with the use of the Mencian analogical model as a decision procedure.

How reflection, reasoning, feeling and desire interweave and mutually influence one another in the *Mencius*

Mencius' use of analogy interestingly also involves using it to extend how we *feel and desire* in past cases to present cases. In trying to persuade King Xuan that he is capable of becoming a true king who can bring peace to his people, Mencius gets him to remember when he spared an ox being led to ritual slaughter. Though the king had been uncertain of his motivation for sparing the ox, Mencius' query prompts him to remember the ox's trembling, which reminded him of an innocent man being led to execution. He tells Mencius that he now feels a stirring in his heart as he remembers what he felt. While the king is re-experiencing compassion, Mencius reminds the king of his duty to his people, to spare them from suffering also. It was probably not Mencius' primary intention to bring home that the king had a duty to his people. He was trying to make the king feel his duty and to do something about it. In contemporary terms, Mencius was trying to turn the King's cognition of his duty from "cool" to "hot." The conclusion of the analogy Mencius was trying to deploy was not merely a judgment but a felt judgment.

Therefore, the *Mencius* concerns both believing and feeling from relevant similarities. By getting the king to re-feel the compassion he felt for the ox at the same time Mencius is reminding him of his people's suffering, Mencius is attempting to enlarge the scope of the king's compassion. Insight can enlarge feeling. At the same time, insight can be made motivationally efficacious through becoming affectively charged. This is an instance of the characteristic interweaving and mutual influence of reflection, reasoning and feeling that is one of the most distinctive and interesting features of Confucian practical reasoning.

This conception of how emotion can be transformed and extended through affect-laden reflection is a corrective to two common and opposing tendencies in construing the relationship between reflection and emotion. One tendency is to think of emotion as a threatening distraction from the kind of dispassionate and objective thinking that ideally guides human conduct. The other tendency is to think of emotion as the ruling force and reason not more than its occasional instrument. In the Confucian

construal, reflection and emotion interact and interweave: through its marriage with feeling, reflection becomes motivationally effective at the same time that it makes the feeling intelligent.

Mencius on exercising judgment when values conflict

Stories told about Shun the sage-king in the *Mencius* feature him as an exemplar whose decisions model *quan*, weighing or discretion, when important ethical considerations conflict. One story about Shun told in the *Mencius* exemplifies the complexity of applying a value such as filiality. It starts with the time when Shun wanted to marry. He knew that his parents would refuse him permission to marry if he were to ask, so he got married without telling them. Mencius approves, explaining that if Shun had let his parents deny him the greatest of human relationships, he would have become bitter toward his parents. Mencius' explanation of this decision is notable for appealing not to the need to override filiality but an interpretation of what action filiality supports when parents are motivated by unreasonable ill will toward a child. It supports not letting parents deny an urgent interest of the child. It is not just for the sake of the child but also for the sake of the filial relationship. It is a common stereotype that Confucianism subordinates the individual to the group, but in this case, Shun's exercise of *quan* (as told by Mencius) is based on the insight that the relationship between the individual and the family is not one of subordination but one of interdependence. To deny one's most urgent interests for the sake of the group is not only to do oneself harm but also in the long term to do no good for the group.

Another story conveys Mencius' interpretation of what Shun would have done had his father murdered another man (7A35). Mencius says that Shun would not have interfered with the Minister of Crime's arrest of his father but then would have abdicated and fled with his father. To appreciate what Mencius said one should compare it to *Analects* 13.18, where Confucius says that a son does not inform the authorities of his father's crime of stealing a sheep but that fathers and sons cover up for each other (alternatively, the language of the original text might be interpreted as saying that fathers and sons do not disclose each other's crimes). Mencius has Shun responding in more complex fashion to his father's crime. Shun could have prevented the arrest of his father, but Mencius pictures Shun as honoring his duty as king. On the other hand, he pictures Shun as removing himself from that role in order to perform his duty as a filial son. Whether or not one agrees with Mencius' particular resolution of Shun's dilemma, it illustrates one of the primary methods of *quan*, which is to seek to reconcile conflicting values as best one can. In this case, it is done over time through performing different actions that honor the values at stake. Trying to reconcile values

involves the exercise of creativity in conceiving of what actions or set of actions might accomplish that task.

Consider by way of contrast the contemporary dilemmas in which one is presented with the choice of stopping a runaway trolley car from hitting and killing five innocent people by doing something that would result in the death of an innocent person. Such dilemmas are presented with pre-defined options from which one must choose and are used to discover the structure of one's moral reasoning (whether it is consequentialist or deontological, for example). Sometimes hypothetical cases are presented to support one type of normative view over another, such as the case of killing and taking the organs of one innocent healthy person to save the lives of five innocent people in need of organ transplants. These uses of cases are perfectly legitimate. But rarely in the contemporary literature is a case presented to illustrate that actual moral problems often do not present us with pre-defined options and that much of the task in moral life is to imaginatively generate options that might resolve or at least mitigate conflicts of values. Mencius' story also illustrates how learning from an exemplar might contribute to such a decision process: one imagines what the exemplar would do in this or that situation. Such imagining does not automatically provide an answer; it rather focuses attention on the salient features of the situation from a perspective formed by concerns that one imagines the exemplar would have and sets the mind to the task of devising a course of action that best answers to those features.

Practical reasoning and Zhuangzi's intuitive activity

It may not appear that the Daoist text ☒☒ *Zhuangzi* (2017) is a promising source for insight into practical reasoning. The text portrays human beings as presumptuous "know-it-alls" who place far too much confidence in reasoning and argument. However, our redeeming quality is our potential to shed our arrogant dispositions and place ourselves in more productive relationship with other things in Nature, both human and nonhuman. To get us to shed our arrogant dispositions, the text applies a constructive skepticism to whatever it is that human beings claim to know. To point towards a more constructive relationship with Nature, it articulates the stance of being a mirror to the world.

Skeptical questioning of what people claim to know is not dogmatic but therapeutic. The aim is to open us to new possibilities of what might become new knowledge and insight, though new discoveries are never immune to questioning in turn, especially when they calcify into received wisdom. In the ☒☒☒ *Xiao yao you* chapter ("Going Rambling without a Destination," *Zhuangzi* 2017) a story is told about Huizi and the gourds. Huizi, Zhuangzi's friend and philosophical sparring partner, grew some seeds that turned into huge gourds. When he tried to put them to some use, he found that they were not sturdy enough to be water containers and too big to be ladles. Huizi gives up, smashing them to pieces. Zhuangzi chastises his friend for not thinking of how he could have lashed the gourds together to make a raft to go floating about on lakes and rivers. That he couldn't see that means his mind is full of underbrush.[1] The underbrush was Huizi's fixed idea of using the gourds to *hold* water, which obscured the use of floating *upon* the water.

Concepts and language enable us to selectively abstract from the potentially overwhelming flow of experience and to focus on resources in the world that we use to satisfy our needs. But we tend to rely on what has worked for us in the past, even if the circumstances have significantly changed. We tend to focus on what we already want instead of what we might find of great value if we were only open to encountering it. Certain things appear into the foreground of perception and push other potentially valuable things into the blurry background. The *Zhuangzi* seeks to unseat conceptions of the world that have become so entrenched that they occlude our vision.

[1] The word translated here as "underbrush" is ☒ *peng*, one meaning of which is a bushy plant, the stems of which can get entangled with one another.

The *Zhuangzi* takes a delight in unseating such conceptions that is akin to the delight that good scientists take in hearing that the received theory has received a jolt from the latest experimental results. Zhuangzi chastises his friend for not clearing his mind of the underbrush of fixed conceptions of how the gourds are to be used, implying that the underbrush can be cleared, at least some of it. More generally, the implication is that we can become aware of how our entrenched concepts and names for things limit us and thus prompt us to try to broaden our current perspectives.

The skeptical stance also sets up the recommendation to be a mirror. In the "☒☒☒ *Ying diwang*" chapter ("The Proper Way for Emperors and Kings"), being a mirror is explained as taking in and reflecting more of what there is in the world, not to let our prior commitments obscure what doesn't fit with them: "Perfected people use their minds like mirrors, not welcoming things as they come or escorting them as they go. They respond without keeping, so they can conquer without harm" (Kjellberg 2001: 242–243). The Z*huangzi*'s sense of being a mirror is different from a crude and literal conception of what it is to be a mirror. It is not transparently, passively depicting the way things are. It is *responding* to things without keeping. Responsiveness that is appropriate but does not keep is an achievement, a skillful activity, and this is why the skill stories in the *Zhuangzi* are relevant to understanding what it is truly like to be a mirror in the Zhuangist sense.

In the "☒☒☒ *Yang sheng zhu*" ("Nurturing Life") chapter (*Zhuangzi* 2017), Cook Ding says that when he first began cutting up oxen, he did not see anything but oxen, but after three years, he couldn't see the whole ox. Now he has so mastered the art that he performs his task in dance-like fashion and does not need to look with his eyes when moving the knife through an ox's joints and spaces. He encounters them with "spirit" (☒ *shen*). He relies "on the Heavenly patterns (☒☒ *tianli*), strikes in the big gaps" and is "guided by the large fissures." This passage has been taken to suggest that the *Zhuangzi* is advocating a kind of inexplicable, mystical access to the natural patterns of the world that enables superlatively effective action. However, the passage, like many others in the text, has multiple levels of meaning, one of which conveys what it feels like to have attained and to perform real-life intuitive skills. Such skills are not in principle inexplicable or accessible only to the mystic but can be built from abilities that are cultivated.

Seeing "nothing but oxen" at the beginning of learning one's craft can be taken to suggest the experience of beginners in a craft – needing to pay attention all at once to what can easily seem to be an overwhelming number of aspects of the skill activity and hence having to self-consciously direct oneself in performing all of it in an exceedingly clumsy way. "Seeing with one's eyes" in the apprentice stage is perhaps a metaphor for having to self-consciously direct oneself in performing all the prescribed moves. Later stages at which one acquires real skill involve mastery through repetition of the basic moves (we might call them "subroutines" of the skill activity) so that one can focus one's attention on doing the things that confer excellence on what one is doing.

"Encountering with spirit" suggests the sense of having acquired a "feel" for what one has to be doing in the moment that does not typically involve thinking about what one is doing but rather the feeling of being totally absorbed in what one is doing with no awareness of a self distinct from the activity. A related feature of skill stories such as Cook Ding's is the embodied nature of the activity. A current trend in cognitive psychology and philosophy of mind has supported the conclusion that much of our perception and action in the world is not well understood by conceiving of our brains as information processors that construct representations of the world based on sensory information and then formulate plans of action based on these representations. Rather, much perception and action flow from the whole of a person's body interacting with the environment. The philosopher of mind Sean Gallagher (2006) has written about the way in which embodiment shapes our minds through our "body schemas," the sensory-motor capacities that give us a sense of our bodies in space.

When we close our eyes and raise an arm, our body schemas enable us to know exactly where our arm is located. When we perceive and act, our body schemas become engaged with features of the environment we can use to accomplish our task. Acquiring skill at using these features means acquiring them as parts of our body schemas. Learning to ride a bicycle is learning how to make it an extension of our bodies. Just as learning to walk is coordinating our movements with the proprioceptive feel of moving forward and maintaining balance, we learn to coordinate our movements with the proprioceptive feel of our bodies on the bicycle as we move forward and maintain balance. Our subjective experience of skillfully using such things in the environment is not of using our minds to direct our bodies in manipulating them but closer to our unself-conscious actions as embodied creatures such as walking or running on our feet. The feeling is one of having enhanced embodied selves making features of the environment extensions of our bodies.

Consider again Cook Ding's feats of cutting. He has made his knife part of his body schema in such a way that he cuts in the unself-conscious way one can ride a bicycle. However, there is a feature of the Cook Ding story indicating that his skill activity is not totally an automated process unmediated by conscious conceptualization about goals, methods and self-consciousness. When Cook Ding comes to a difficult spot in the ox, he says he must pause and gather himself to make a careful effort. This moment in the cook's description of his own activity implies that experiential immersion in one's activity is never absolute and total nor the feeling of flowing effortlessness. The capability for taking self-conscious control of one's activity gets triggered, perhaps when there is an interruption in the feel of flowing effortlessness. A more complex picture of intuitive activity emerges, one that allows conscious thought, goal-directness and being guided by thoughts of method to interact and/or alternate with automatic non-conscious doing.

The ☒☒☒ *Renjianshi* ("In the Human World," *Zhuangzi* 2017) chapter presents an activity that is skillful in another way in that it involves working with people and the attempt to do something worthwhile in the social world. Yan Hui tells Confucius that

he wants to go to a state and reform a young and reckless ruler. Confucius advises Yan Hui to not bring into the situation any preconceived plans for accomplishing his goal. Just as the cook must develop his ability to intuitively find with his knife the great hollows and cavities of the ox, so Yan Hui, if he puts aside his ambitions and his pre-conceived plans for changing the young ruler (what Confucius in this chapter calls fasting the mind), has a chance of navigating skillfully his way inside the psyche of this ruler and turning him towards a better course.

Being a mirror in this case means that he is prepared to determine how to do that in the process of dealing with the ruler, having set aside all preconceptions of how to do that. Yan Hui must listen not with his mind but with his ☒ *qi* or bodily vital energies. While the mind tends to impose its preconceived names and conceptions of how things and people are, the body's energies are responsive to how things on the outside are. Here there is resonance with Antonio Damasio's (1994) seminal theory of somatic markers. In this theory, subjective feeling is constituted by the mental states arising from the neural representation of various changes occurring within the chemical landscape of the body. These somatic changes are responses to a precipitating event and mark the event positively or negatively so that it stands out as an object for choice. In the ancient Chinese conception of the person, our interactions with others leave their marks on our *qi*. Our bodily energies record the positive and negative impressions others make on us. There is knowledge of who they are that is embedded in us, waiting to be taken advantage of if we are open to their input.

In fact, humans do have impressive abilities to read one another's nonverbal language of physical postures, gestures and facial expressions, and this reading often sits below the level of conscious awareness. Participants in one study were able to identify facial expressions they glimpsed for only five milliseconds (Rosenthal et al 1979). In another study, participants were able to order fourteen photographs of the temporal sequence of an emotion unfolding over the course of less than one second (Edwards 1998). A third study reported that participants were able to detect minute violations of the basic proportions of the human face (Lewicki 1986). They felt something was wrong with the faces, but none of them were able to identify what was wrong. Such skill in "nonverbal decoding" improves from early childhood through early adulthood (Cohen et al 1990; Dimitrovsky 1964; Hamilton 1973; Rosenthal 1979). As Matthew Lieberman observes, "the dance of nonverbal communication between people occurs intuitively, and when we get a sense of the other's state of mind as a result of the nonverbal cues the other has presented, we often have nothing other than our intuition to justify our inferences" (2000: 123).

Interestingly, there is some evidence to suggest that deliberate, conscious learning can interfere with the kind of "implicit" or nonconscious learning under discussion (Lieberman 2000: 121). Thus the relation between conscious activity guided by thoughts of method and intuitive non-conscious activity is complex. Thought focused on method can lead to practice aimed at acquiring skill so that much of what is initially guided consciously can subsequently become automatic. Conscious thought can also

take over when unconscious activity hits problematic spot. But at other times, it can be better to suspend conscious thought in favor of non-conscious processing, because the two can interfere with one another. This returns us to the Zhuangist point that our names, concepts and existing likes and dislikes can filter from our experience much of what could be potentially valuable to us. Our penchant for naming things and pinning them down with our concepts can cut off a more informative receptivity to what they are.

The kind of automatic processing that is depicted in the *Zhuangzi* is different from the kind that involves "heuristics and biases" that allow us to respond in quick and dirty ways to events and things in the world that are critical to our survival and reproduction but may especially in modern circumstances lead to error (Kahneman and Tversky).[2] The *Zhuangzi* draws our attention to capabilities of picking up on very complex patterns that get revealed after repeated experiences of the relevant sequence of events. These might fairly be called nonconscious inferential processes and therefore part of the structure of the practical reasoning of human beings, which under this portrayal (and also under the Confucians' portrayal) turns out to be more complex and interesting than is revealed by the standard philosophical models that tend to be conscious and rule guided. As Isenman suggests, some of our most useful intuitions are produced not through a simplifying heuristic applied to a thin slice of experience, but non-conscious processing that integrates multiple clues into a meaningful complex pattern that may be "too multidimensional and interwoven for the conscious mind, with its ability to hold a very limited amount of information at the same time, to comprehend, never mind articulate" (2013: 160).

If there is sometimes competition between the automatic processing of experience and explicit, deliberate analysis, and if relevant experiential patterns are too complex for conscious analysis, it may make sense, as the *Zhuangzi* does, to encourage suspension of the latter in favor of the former. Instead of always trying to analyze and regiment the meaning of our experience with names and concepts, we can promote a receptive attitude toward things and other people, waiting for them to teach us and being receptive to what our bodies and non-conscious levels of mind have to say about what we have been taught.

[2] What gets called "intuition" in the scientific literature often includes conclusions reached on the basis of quick and dirty heuristics, such as stereotypes that people rely upon in reading others' emotions (e.g., stereotypes of emotions associated with people in certain social roles or groups). Ma-Kellams and Lerner (2016) found that empathic intuition of others' emotions was less accurate than systematic thinking about others such as perspective-taking, but since their conception of intuition includes the use of heuristic stereotypes, they are not talking about the kind of intuition the *Zhuangzi* is concerned with.

Synthesizing Confucian and Daoist practical reasoning

Intuition is not always reliable, of course, and is subject to all kinds of biases. The kind of sensitivity to one's own internal emotional life that Confucianism advocates can help in cultivating a greater awareness of the conditions under which one's intuitions tend to be unreliable. At the same time, the *Zhuangzi* presents a well-taken caution to Confucians who might too readily assume that they grasped the nature of moral excellence and of all the value to be found in the world. Taken together, the early Chinese philosophical tradition presents a conception of practical reasoning that remains relevant and is a corrective to exclusively formal models.

Further reading

Graham, A. C. (2003) *Later Mohist Logic, Ethics and Science*, Hong Kong: Chinese University Press.

Harbsmeier, C. (1993) "Conceptions of Knowledge in Ancient China," in *Epistemological Issues in Classical Chinese Philosophy*, ed. H. Lenk and G. Paul, pp. 11–30, Albany, NY: SUNY Press. Concise and informative survey of conceptions of knowledge in the classical tradition.

Lau, D. C. (1970) "On Mencius' Use of the Method of Analogy in Argument," in *Mencius*, trans. D. C. Lau, pp. 235–263, London: Penguin Books. Influential explication of analogical reasoning in Mencius.

Makeham, J. (1994) *Name and Actuality in Early Chinese Thought*, Albany, NY: SUNY Press.

Olberding, A. (2013) *Exemplarism in the Analects: The Good Person Is That*, New York: Routledge. Makes a persuasive case that the epistemology of the *Analects* starts from concrete exemplars, with generalizations and further systematization grounded in exemplars.

References

Analects (2017) Chinese Text Scanned and edited by *Chinese Text Project* http://ctext.org/analects, ed. D. Sturgeon, from *The Chinese Classics*, vol. 1, ed. and trans. J. Legge, Oxford: Oxford University Press, 1861.

Cohen, M., Prather, A., Town, P. and Hynd, G. (1990) "Neurodevelopmental Differences in Emotional Prosody in Normal Children and Children with Left and Right Temporal Lobe Epilepsy," *Brain & Language* 38: 122–134.

Damasio, A. R. (1994) Descartes' Error: Emotion, Reason, and the Human Brain*, New York: Putnam.*

Dancy, J. (1993) *Moral Reasons*, Oxford: Blackwell.

Dimitrovsky, L. (1964) "The Ability to Identify the Emotional Meaning of Vocal Expressions at Successive Age Levels," in *The Communication of Emotional Meaning*, ed. J. R. Davitz, pp. 69–86, New York: McGraw-Hill.

Edwards, K. (1998) "The Face of Time: Temporal cues in Facial Expressions of Emotion," *Psychological Science* 9: 270–276.

Gallagher, S. (2006) *How the Body Shapes the Mind*, Oxford: Oxford University Press.

Hamilton, M. L. (1973) "Imitative Behavior and Expressive Ability in Facial Expressions of Emotions," *Developmental Psychology* 8: 138.

Hamlin, J. K. and Wynn, K. (2011) "Young Infants Prefer Prosocial to Antisocial Others," *Cognitive Development* 26: 30–39.

Hamlin, J. K., Wynn, K. and Bloom, P. (2007) "Social Evaluation by Preverbal Infants," *Nature* 450 (22): 557–560.

———. (2010) "Three-Month-Olds Show a Negativity Bias in Their Social Evaluations," *Developmental Science* 13 (6): 923–929.

Isenman, L. (2013) "Understanding Unconscious Intelligence and Intuition: '*Blink*' and Beyond," *Perspectives in Biology and Medicine* 56 (1): 148–166.

Kjellberg, P. (2001) Translation of Selections from the *Zhuangzi*, in *Readings in Classical Chinese Philosophy*, 2nd ed., ed. P. J. Ivanhoe and B. W. Van Norden, Indianapolis, IN: Hackett Publishing Company.

Lewicki, P. (1986) *Nonconscious Social Information Processing*, New York: Academic Press.

Lieberman, M. D. (2000) "Intuition: A Social Cognitive Neuroscience Approach," *Psychological Bulletin* 126: 109–137.

Ma-Kellams, C. and Lerner, J. (2016) "Trust Your Gut or Think Carefully? Examining Whether an Intuitive, Versus a Systematic, Mode of Thought Produces Greater Empathic Accuracy," *Journal of Personality and Social Psychology* 111 (5): 674–685.

Mengzi (Mencius). (2017) Chinese Text Scanned and edited by *Chinese Text Project* http://ctext.org/mengzi, ed. D. Sturgeon, from *The Works of Mencius*, ed and trans. James Legge, Oxford: Clarendon, 1985.

McNaughton, D. A. (1988) *Moral Vision*, Oxford: Blackwell.

Olberding, A. (2008) "Dreaming of the Duke of Zhou," *Journal of Chinese Philosophy* 35 (4): 625–639.

———. *(2013)* Exemplarism in the Analects: The Good Person Is That*, New York: Routledge.*

Rosenthal, R., Hall, J. A., DiMatteo, M. R., Rogers, P. L. and Archer, D. (1979) *Sensitivity to Nonverbal Communication: The PONS Test*, Baltimore: John Hopkins University Press.

Warneken, F. and Tomasello, M. (2007) "Helping and Cooperation at 14 Months of Age," *Infancy* 11 (3): 271–294.

Wong, D. B. (1991) "'Is There a Distinction Between Reason and Emotion in Mencius?' and a Reply to a Commentary by Craig Ihara," *Philosophy East and West* 41: 31–58.

———. (2002) "Reasons and Analogical Reasoning in Mengzi," in *Essays on the Moral Philosophy of Mengzi*, eds. X. Liu and P. J. Ivanhoe, 187–220, Indianapolis, IN: Hackett Publishing Company.

———. (2015) "Growing Virtue: The Theory and Science of Developing Compassion from a Mencian Perspective," in *The Philosophical Challenge from China*, ed. B. Bruya, Cambridge, MA: MIT Press.

Zhuangzi. (2017) Chinese Text Scanned and edited by *Chinese Text Project* http://ctext.org/zhuangzi, ed. D. Sturgeon, from *The Writings of Chuang Tzu*, ed. and trans. J. Legge, Oxford: Oxford University Press, 1891.

8. Aristotle on deliberation[1]

[1] This chapter was presented at Auburn University's 2016 conference, "Aristotle and Kant in Conversation," at UCLA and at the Humboldt University, Berlin. I am grateful to those audiences for their feedback, as well as to Stephen Menn and Susan Sauvé Meyer.

Agnes Callard

Described schematically, Aristotle's theory of deliberation (*bouleusis*) is immediately familiar as a theory of what we,[2] too, would call deliberation: a conscious, rational mental processes deployed by an agent in order to solve practical problems. He thinks, as we do, that deliberation is a form of thought that takes time, proceeds systematically rather than haphazardly, and ends by putting the agent in a position to choose rationally[3]. Deliberation, for Aristotle as for us, is thought that answers the question, "What should I do?"

We take deliberation to be the most direct manifestation of someone's mastery of practical rationality, and Aristotle does as well: he makes excellence at deliberation the characteristic virtue of the ideally rational agent or *phronimos* (*Nicomachean Ethics* [NE], 1141b10). The phronimos is a good person all around, but the virtue that marks him *as* phronimos is the fact that he is good at deliberating. Aristotle is also attuned, as we are, to the fact that deliberation does not represent the whole of practical rationality. He acknowledges that much human action[4] – action that may be courageous, generous or just – is produced without recourse to deliberation.[5] Sometimes we know what to do immediately, and we simply act without having to think about what we should do. Like us, Aristotle thinks that we will find paradigm instances of deliberation in hard cases, which is to say, when it is not obvious what a person should do.[6]

In its broad outlines, then, Aristotle's conception of the function of deliberation is akin to our own. This makes it all the more surprising that his conception of the mental operations of the deliberator is not. We understand deliberation as a process by which someone evaluates an action she is considering performing. This evaluation takes one of two forms. The first form is comparison: an agent deliberates in order

[2] I use a contrast between "our" approach to deliberation and the one I ascribe to Aristotle to ease exposition throughout this chapter, with apologies to those readers whose sympathies and intuitions align more closely with the Aristotelian approach than with the evaluative one I ascribe to "us."

[3] John McDowell 1979 has taken Aristotle's account of deliberation to be less a theory of a certain kind of thought process than an account of the rational structure inherent in action; see also Cooper 1986, pp. 9–10; here I follow Price 2011; Segvic (2009, pp. 149–153), who conceive of deliberation as a process of occurrent thought, "a plausible sequence of mental acts" (Price p. 155, describing both his view and Broadie's). See also Broadie 1991, n.11, pp. 118–119. As Broadie emphasizes, it may nonetheless be possible to use what Aristotle has to say about deliberation to shed light on the rational structure of non-deliberated action. Moreover, it is important to keep in mind Cooper's (pp. 7–8) point that the deliberation may be performed well in advance of the action.

[4] Moreover, Aristotle is careful, as we sometimes are not, to make room for rational reaction in addition to rational action; see Kosman 1980 for a discussion of the rationality of feeling.

[5] See especially his discussion of the value of sudden (as opposed to calculated) acts of courage at NE III.8, 1117a17–22. For discussion of the question as to how it is possible for such action to be virtuous, given the connection between deliberation, choice and virtue, see Segvic 2009, pp. 157–159; Broadie 1991, pp. 78–82.

[6] We deliberate more in those crafts that have been "less precisely worked out" and more generally on matters where we are "in two minds" and where "it is unclear how they will in fact fall out" (NE III.3, 1112b5–9).

to select the best among a set of available options. The second is one in which she assesses a single action by way of some (usually moral) rule, having committed to performing only actions that pass the test in question. Evaluations of the first kind usually include some consideration of what consequences may arise from the proposed action, where the deliberative import of any potential consequence is tempered by the agent's understanding of the likelihood of that consequence. For instance, an agent might aim to perform the action that is likeliest to maximize the satisfaction of her preferences. The second form of evaluation is associated with Immanuel Kant, and its deployment typically involves less concern with the consequences of an action than with the internal or formal qualities of the action itself. Thus Kant famously concludes that one ought not lie to the murderer at the door, because the action of lying is one that the agent cannot will to be a universal law of nature.[7] These two forms of evaluation do not necessarily compete with one another: the Kantian deliberator may well fall back on comparative assessment when, for instance, she ascertains that multiple proposed actions are sanctioned by her 'categorical imperative.'[8]

Aristotle does not have an evaluative model of deliberation.[9] Aristotle thinks that the deliberator begins with a goal or target or end, the realization of which is both desirable and difficult: she cannot immediately see how to bring it about. She reasons backwards from this end, working out the process by which she might bring it into being. Drawing the end into the sphere of her own agency, she eventually hits upon something she sees she can do. This action then becomes the object of her choice. Aristotle's agent *evaluates* neither the goal with which she begins her deliberation nor the action in which her deliberation ends. Instead, her deliberation consists in the mental activity of *deriving the action from the goal*.

In this chapter, I will first explain how Aristotelian deliberation works, then contrast it with the evaluative deliberation with which we are more familiar. A familiar worry will arise: given that Aristotelian deliberation is an inquiry into the means for a given end, does it amount to Humean instrumental reasoning? I explain why it doesn't while acknowledging that Aristotle agrees with Hume about the impossibility of deliberating about ends. Finally, I'll offer an overview of the difference between the evaluative and the Aristotelian approach to deliberation.

---

[7] AK4:436–7.

[8] The function of both the Kantian and the comparative model is broader than that of deliberation, since we will not always need to occurrently think through the issues in question; my interest, however, is restricted to the deliberative function of each.

[9] There has been a tradition of trying to squeeze Aristotle into the modern mold, for an overview of which, see Nielsen 2011. A number of more recent commentators, most notably Nielsen herself, have insisted that this will not work. See Normore 1998, who observes that "it is not crucial to human choice (prohairesis) that the agent be confronted with several means to an end" (p. 25). See also Wiggins 1980, p. 232.

I A geometrical model for deliberation

What is it to derive an action from a goal? Aristotle suggests that a similar form of derivation occurs in geometry:

> For the person who deliberates seems to investigate and to analyze in the way we have said, as if with a diagram (and while not all investigation appears to be deliberation, as e.g. mathematical investigations are not, all deliberation is investigation); and what is last in the analysis seems to be first in the process of things' coming about.
>
> *NE* III.3, 1112b20–4[1]

Aristotle is here describing a mathematical operation of analysis,[2] which is a tool used by geometers in navigating tricky construction problems. Suppose I am told 'construct a square inscribed in a circle.' If I know how to do this, I will begin by drawing the lines from which the square will arise. That is called the synthesis. But suppose I do not know how to do this. How am I to decide which lines to draw? One thing I could do is begin by analysis, which entails assuming that I have before me the very figure I have been assigned to construct – a square inside of a circle (see _Figure 8.1]]) – and reasoning backwards to the way it was been produced. This form of reasoning involves adding elements to the diagram that can be constructed from it or noticing relationships among its parts that are, in turn, geometrically determined by the assumption of the completed construction.

Here is an illustration of how one might use analysis to solve the problem I just described.

In the first step (A1), we have assumed a square inside of a circle. In A2, it occurs to me that if I had the square's diagonals, I could use them to draw the square. I then notice that the diagonals are at a right angle to one another (A3) and that they intersect at the center of the circle (A4). This focuses my attention on the project of drawing just one of the diagonals/diameters – because, given my ability to draw a perpendicular bisector, I know I can use it to draw the other. At this point, I have

[1] Translations of the *NE* are from Rowe 2002; other translations of Aristotle are from Barnes 1984.

[2] This account of analysis draws heavily on Menn 2002, who discusses the distinction (due to Pappus of Alexandria) between the "problematic" analysis I describe here and "theoretic" analysis. The

come to see that if I can draw the diameter of a circle, I can draw a square inside that circle. But how to draw the diameter? It occurs to me that the diameter of the circle lies perpendicular to the midpoint of any chord of the circle. And now I see that if I draw a chord to the circle, I can construct a diameter (A5). For I can draw the perpendicular bisector of the chord. At this point, something clicks and I have an "aha!" moment: I see that the chord on which the rest of the figure depends is one that I can simply draw on the basis of the circle alone. Elsewhere, Aristotle describes this final moment of the analysis – both geometrical and deliberative – as the work of a form of insight (*nous*, NE VI.8, 1142a25). From this point, I can reverse the procedure and perform the (synthetic) construction (S1–5).

Let us make some observations about the various forms of competence this reasoning called for. The geometer I was imagining had in her possession some geometrical knowledge. She knew that the line perpendicular to a chord runs through the center of the circle; that the center of a square is the center of the circle in which it is inscribed; that the angles determined by the diagonals of a square are right angles. She also had skills, for instance, the ability to construct both a perpendicular bisector of a line segment and the line perpendicular to a line at a given point on that line. She also evidently possessed some expertise in and familiarity with methods of construction: she saw drawing the diagonals in A2 or the chord in A5 as a useful moves in the construction. It takes some experience in geometry to see that the diagonal or the chord are lines from which the requisite figure – the square and the diameter, respectively – could arise.

Analysis offers someone a way to exploit her geometrical knowledge, skills and experience to derive the first step of her construction – the drawing of the chord – from a representation of the figure she is trying to construct. She operates on that figure both by adding (lines, arcs, points) to it and by noticing relationships that obtain as a result of these additions.[3] In the analysis, she is figuring out how to bring a certain figure into being, and in the synthesis, she actually brings it into being. We can observe this difference in the two sets of diagrams: construction marks (e.g., the arcs in S3 and S4) are only present in the synthesis.

How can we translate this account of geometrical reasoning into a model of practical thinking? In the *Metaphysic*s, Aristotle shows us how something like analysis appears in the deliberation of a craftsman:

> What is healthy comes into being when the producer has had the following sort of thought: since health is this, then if something is to be healthy, it

latter is concerned with proofs of given propositions rather than construction problems (p. 199).

[3] It is not quite right to say, as Nielsen 2011 does, that "the geometrician first identifies its smallest parts and then constructs these. Ultimately, she is able to construct the entire complex figure by breaking it down and constructing its simple constituents step by step" (ibid., p. 401). Quite often, what one needs to construct will not be, except in a very extended sense of the term, a "part" of the original picture (such as the chord in my example or the diagonal line in the example at *Meno* 82–85).

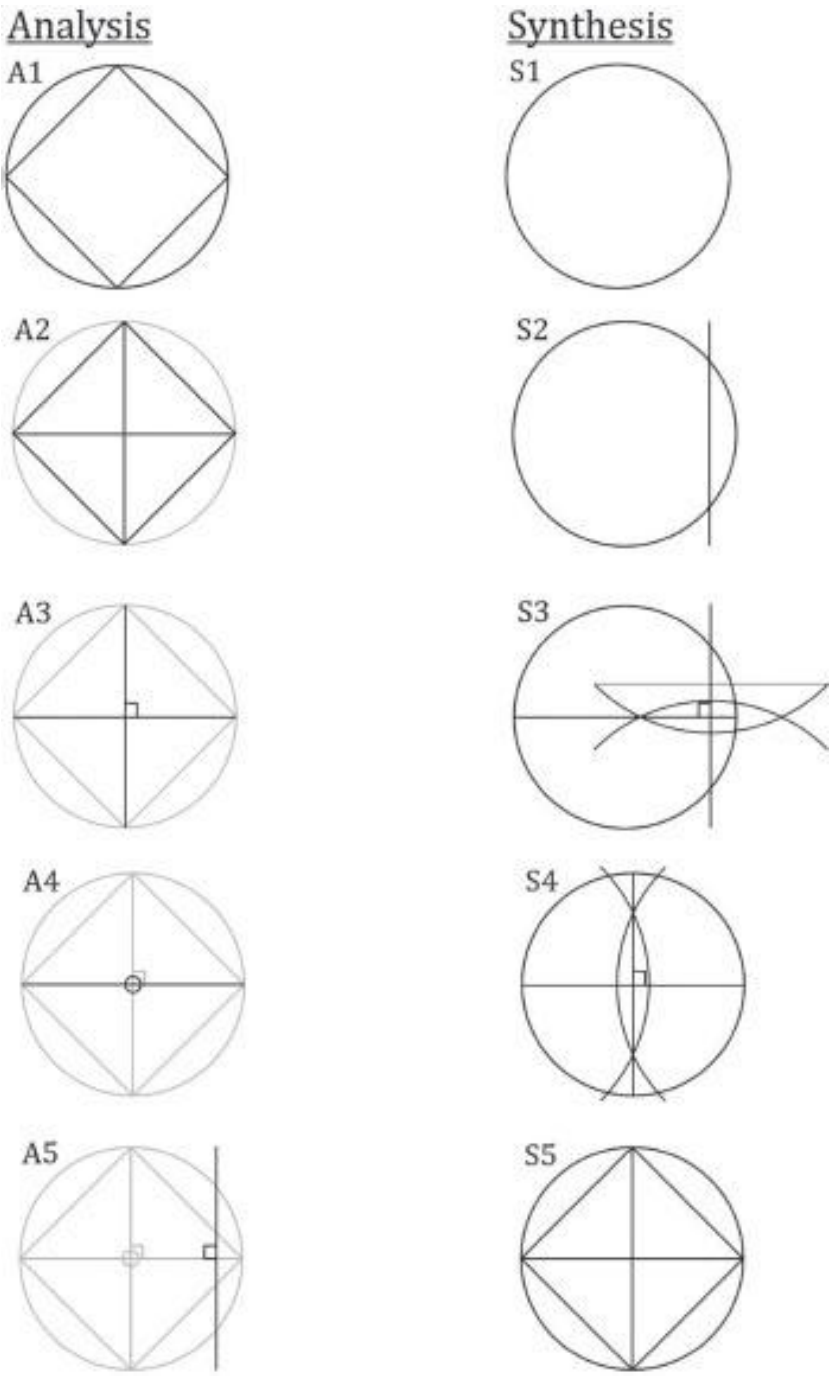


*Figure 8.1*Inscribing a square inside a circle

> must have this (for instance, a uniform condition of the body), and if it is to have this, it must have heat. This is how he thinks at each stage, until he leads the process back to the last thing, which is what he can produce himself; and then the motion from here on toward health is called a production... . Production is the motion that proceeds from the last stage in thinking. Each of the other things – those in between – comes to be in the same way. I mean, for instance, that if this [body] is to be healthy, its bodily condition must be made uniform. What then, is it to be made uniform? This. [The body] will have this if it is warmed. What is it to be warmed? This. But this is potentially present. And now he has reached what is up to himself.
>
> (*Meta.* Zeta, 1032b6–22)

The doctor begins by thinking of his goal, health, and then reflecting on what health is. He comes to some conclusion about what health is, and reasons that *that* (Aristotle doesn't tell us what), calls for a uniform condition of the body. He then moves to the thought that heating produces uniformity. But applying heat is something he can do – it is "potentially present" not only in the analysis but in the synthesis, just as that the chord is potentially present not only in Figure 8.1 A1–5 but also in S1. The doctor's reasoning involves both instrumental 'additions' to his picture of health, such as the idea that heating produces uniformity, as well as moments where he is examining the relations between its parts rather than adding to it. The latter forms of reasoning (e.g., "since health is this") parallel the geometer's observing that, for instance, the diagonals intersect at right angles in A3.[4]

One striking feature of Aristotelian deliberation that emerges from this passage is the fact that it begins with a fixed, given end. In order to do the work of deriving the action, one must hold that end fixed. The end of health sets the problem to be solved, just as the project of constructing the square in the circle does in the geometrical case. This feature – the fixity of the end – is not just a feature of mathematical reasoning or of craft-deliberation, but holds also in case of properly ethical deliberation. This becomes clear in the long discussion of deliberation in *Eudemian Ethics* (EE) II.10, from which I excerpt two key passages:

> But the cause or object will come first, e.g. wealth, pleasure, or anything else of the sort that happens to be our object. For the man deliberating deliberates if he has considered, from the point of view of the end, what

[4] We thus have a textual basis for the distinction that is sometimes drawn between constituent means and productive means. One can reason about what the end amounts to, or one can reason about what will give rise to it. This distinction has taken on a special significance extending beyond the study of Aristotle, since constitutive reasoning promises to broaden our conception of instrumental reasoning. See Wiggins 1980, p. 224; Cooper 1986, p. 22; Nussbaum, 2001, p. 297; Sorabji 1980, p. 202. These authors have rightly emphasized the significance, for Aristotle, of the fact that deliberation involves this

> conduces to bringing the end within his own action, or what he at present can do towards the object.
>
> (EE 1227a13–18)
>
> Now about the end no one deliberates (this being fixed for all), but about that which tends to it – whether this or that tends to it, and – supposing this or that resolved on – how it is to be brought about. All consider this till they have brought the beginning of the process to a point in their own power.
>
> (EE 1226b10–13)

Notice Aristotle's emphasis on the fixity of the end over the course of the agent's deliberation: Aristotle is imagining a person who has as his end wealth or pleasure and then derives from that end some action he can perform.[5] Compare this to the evaluative deliberation with which we are more familiar. There, what is fixed is the set of potential actions that stand before the agent as options: in order to evaluate whether an action is best overall, or satisfies the moral law, we must know *which* action or actions we are talking about. In order to form what Davidson (1980, p. 39) calls an "all things considered" judgment as to what one ought to do, one needn't in fact consider "all things." But one does need to close off the space of options, effectively counting whatever set of options one is considering *as* all the options that there are.[6] Thus evaluative deliberation operates by fixing the options. Nothing prevents the agent engaged in such deliberation from having in view multiple distinct ends which these potential actions might serve.[7]

Aristotle does not understand deliberation as a process of trying to figure out whether a given option is an acceptable option or the best among a certain set of options. Rather, one deliberates about how to bring about an end such as pleasure or money. Aristotle's deliberator isn't considering some candidate mode of financial gain or even assuming that there is such a mode. He says that sometimes what deliberation reveals there is *no* option you can take (1112b25). Then you give up. The work of deliberation is to find the analytic path to a single option, rather than to select between given options.

---

constitutive element, though they have not always acknowledged the way in which the constitutive and the productive forms of reasoning work together in Aristotle's examples. It is hard to see how we could have the one without the other in a successful case of (geometrical or ethical) analysis.

[5] For discussion of why the role of the end cannot be filled by some formal object such as "the mere unrestricted good, the formal end of practical deliberation," see Broadie 1991, p. 233, and n. 51 p. 262.

[6] This is not to say that new options cannot arise in the course of deliberation, but that if they do, the finality of the judgment is thereby undermined, and one must re-fix the set of comparanda to include the new item.

[7] Proponents of the evaluative view may regard the end as fixed but as formal or indeterminate: happiness, the good, whatever best satisfies my preferences (see n. Error! Bookmark not defined.). Some utilitarians may regard the end as both fixed and determinate, for example, pleasure. Thus the clearest

Aristotle allows that such a search may include comparative reasoning. Consider this passage from NE III.3:

> Having set the end they consider how and by what means it is to be attained; and if it seems to be produced by several means they consider by which it is most easily and most nobly produced, while if it is achieved by one only they consider how it will be achieved by this and by what means *this* will be achieved, till they come to the first cause, which in the order of discovery is last.
>
> (NE III.3, 1112b15–20)

Aristotle anticipates the possible need for comparison arising during deliberation. But rather than identifying the work of comparing with the work of deliberation, he sees comparison as a (possible) step in the course of deliberation. Moreover, what one compares are not options but rather means that will need to be further determined ("and by what means *this* will be achieved") before they can be chosen. Notice also that while Aristotle acknowledges not only that there may be multiple means to a given end, and even that there may be multiple criteria (ease and nobility) for selecting among means, he does not offer any advice as to how to weigh these against one another. He does not seem disturbed by the fact that the easiest way to do something is not usually the most noble.

We can consider the analogous situation in the case of the geometric construction: suppose there is a particularly beautiful way to do a proof, but it will take me more time because, for example, I will need to go get my compass from the drawer. Aristotle seems to be saying: choose the more elegant one, or the quicker one, whichever! Multiple ways of achieving an end seem to strike Aristotle less as a source of profound deliberative challenge than as an embarrassment of riches. Comparison is an occasional wrinkle of Aristotelian deliberation; by contrast, in the evaluative deliberation familiar to us, comparison typically constitutes the agent's entire deliberative work. For Aristotle, the chief deliberative work is that of finding the means.

This fact about deliberation fits into Aristotle's bigger moral-psychological picture: he posits a division of labor between the intellectual part of the soul (to logistikon), which has the function of deliberating (as well as engaging in theoretical reasoning), and the affective part of the soul, in virtue of which one has feelings and desires (NE I.13). Famously and problematically, Aristotle allots to the affective part the job of providing the ends in the service of which we deliberate. Desire or affect or feeling grasps the end, while deliberation discovers the means (NEVI.12–13).[8] Aristotle, like Hume, denies that we can deliberate about ends. Does that mean that Aristotle, like Hume, relegates deliberation to the discovery of a causal pathway to whatever we

point of contrast with Aristotle will be on the question of the fixity of the options.

[8] For a discussion of the controversy over whether we should (as I do) take these division of labor

happen to desire? Does he hold reason to be impotent? Does he agree with Hume, who says that "Reason is, and ought only to be the slave of the passions" (1739, 2.3.3)?

passages at face value, and a defense of doing so, see Moss 2011.

Backwards reasoning[1]

Hume's theory of motivation allots to reason the role of scouting out a causal link between the subject and her desired object. In his *Treatise of Human Nature*, he says that we "cast our view on every side" in order to seek out

> whatever objects are connected with its original one by the relation of cause and effect. Here then reasoning takes place to discover this relation; and according as our reasoning varies, our actions receive a subsequent variation. But it is evident in this case that the impulse arises not from reason, but is only directed by it.
>
> (1739, 2.3.3)

Hume is explicit that he is describing a form of theoretical reasoning discussed earlier in the *Treatise*, namely probabilistic reasoning about causes and effects. One could just as well engage in this reasoning without any desire for the end and without being in a position to supply the means.

The fact that you could do this kind of reasoning without wanting anything and without having any practical capacities indicates a symmetry between the starting point and the endpoint of Humean practical reasoning. The two points are like dots on a piece of paper you are trying to connect. If there are relatively few ways to realize your goal, and relatively many things you are immediately in a position to do, you might start your reasoning from the endpoint; if you have relatively few immediate options, and there are many ways to get to your goal, it will make more sense to start reasoning from your current standpoint.

If Aristotelian deliberation were just a matter of scouting a link between means and ends, then Aristotle would find it natural to describe us as working forward (from what we conceive as possible for us) at the same time as we work backwards (from what gives rise to the desired result). For we would be engaged in a project of matching the causes we can produce with the effects we desire. But Aristotle states explicitly, by way of the comparison with analysis, that deliberation is unidirectional: it moves backwards from the end to the action.[2]

[1] Thanks to Tom Lockhart for raising a crucial objection that helped me re-think this section.

[2] Thus I disagree with Broadie's (1991) claim that it is the work of deliberation "to convert the agent's particular situation into the elements of a realised good action" (p. 227). For Aristotle is clear that what gets converted is the *end,* not the situation. I grant, of course, that the agent's appreciation of her situation must figure in the conversion process (see note Error! Bookmark not defined.), but I

For Aristotle, deliberation is an asymmetrical mode of reasoning: it systematically transforms wish, a desire for the end, into choice, a desire for the means as a way of getting the end. (Hence choice, *prohairesis*, is a getting [*hairesis*] of one thing – the means – in place of [*pro*] another thing, the end.) The transformation must follow certain rules, which is why it must proceed in the direction it does. In order to see the work these rules are doing, we must spend a moment reflecting on the kind of object that Aristotle takes an end to be.

Suppose, to take an end suggested in our *EE* II.10 passages, that the deliberator is investigating how to acquire wealth. Aristotle would refuse to describe her as someone who simply aims to "bring it about" that she has a lot of money, any more than the geometer simply aims to "bring about" a drawing that resembles the picture in A1. In order to generate a properly geometrical object, her process of generation must respect geometrical procedure. If she were, for instance, to *trace* the drawing of the inscribed square I produced previously, the object she generated would be the wrong kind of square. Instead of constructing a geometrical square, she would have merely drawn a square-shaped mark. Likewise, the ethical reasoner would not be satisfied with bringing about a lot of money in a way that would, for instance, result in her being dead or otherwise unable to use the money.

She doesn't just want money, she wants money insofar as it is something good for her. We could call her object "good wealth," but that would be a bit repetitive, in the way that I could have been accused of being repetitive when I spoke, a minute ago, of constructing the "geometrical square." A real square just *is* a geometrical square, just as, according to Aristotle, wealth just is a good thing: he calls it good absolutely (*haplōs*, see references subsequently). Better to say, by wealth she is referring to an item in the ethical domain, which is to say, some good; just as by the square the geometer is referring to an item in the geometrical domain, which is to say, some ideal figure.

The geometer must take care that each move in the analysis observes geometrical rules, lest his construction devolve into a mere drawing. Analogously, the ethical reasoner must take care to observe ethical rules, lest her deliberation devolve into a mere "bringing about." If she acquires a lot of money at the cost of, say, her life or health, she would not have acquired the ethical object that was her end. This is why Aristotle insists that good deliberation is *not* merely a matter of securing money, but of doing so "as a result of correct reasoning" (VI.9, 1142b16). Aristotle holds that there is a rational procedure for preserving the ethical status of the wealth one acquires that is analogous to the rational procedure for preserving the geometrical status of the square one inscribes. This rational procedure is what he takes deliberation to be.

---

believe the assumption that it takes a conception of her situation as a starting point leads Broadie, without textual basis, to import the evaluative model into her reading of Aristotle. For instance, she glosses deliberating as a matter of "considering alternative possible actions each of which presents itself as loaded with its own set of reasons" (p. 227). Notice how this description immediately makes deliberation a matter of comparison, as opposed to derivation.

Aristotle understands deliberation as having the hypothetical[3] structure analogous to the one found in geometrical analysis: *if* I could construct the diagonals, I would be able to construct the square, and *if* I could construct one diagonal, I could construct the other, and *if* I could construct a chord, I could construct the diagonal. But wait: I *can* construct a chord! And you are off. Every step in the analysis is *a move* in the analysis – which is to say, it is *derived* from the previous step. And this is why deliberation is unidirectional: the agent is deriving her action from her end.

We might still wonder, "Why can't the deliberator reason forwards from where she stands at the same time as she reasons backwards from what she wants?" The answer is that such forwards movement could not be a form of ethical reasoning. The fact that the agent is situated at some place or time is not a fact from which anything rationally follows. It is true that her situation makes it possible for her to do some things and not others, but she has no principle for selecting among these in her reasoning: none of them is rationally necessitated, and thus nothing follows ethically from what resources she has at her disposal. From the fact that an agent has such-and-such powers, it does not follow that activating any of them would be good for her. Likewise, the geometer confronting the empty circle can draw many lines, but none of them would *follow* from anything.

Humean instrumental reasoning is successful insofar as it traces a path from the agent to her end, allowing her to secure the object of desire (e.g., wealth). In Aristotelian reasoning, this is not enough. One must not only secure the object but secure it through a derivational procedure that ensures the end retains its ethical status – its goodness. This is why Aristotle insists that deliberation move backwards. And this insistence, in turn, implies a deeper difference in how Aristotle and Hume understand practical reasoning. In the next section, I will explain how Aristotle's analytic conception of deliberation gives rise to his view that vicious people can't reason well.

Aristotle's account of the rational transformation of wish into choice provides an excellent diagnosis of why, in fairy tales, wishes so often go awry: if your wish "comes true" without your doing the work to make it come true, that means you aren't around to guide the wished-for end, rationally, into existence. Its not surprising, then, that we so often get, as they say, a slip twixt cup and lip – you get to marry the princess, but she's a jerk; you get to live forever, but as a grasshopper; you get a delicious hamburger but, to borrow John Searle's example, it's encased in Lucite. If you let the end out of your hands and entrust it to a genie or magical force, there is no guarantee it will *stay* good. That's the job of practical reasoning. Perhaps we can even say that the basic moral of wish/genie/fairy godmother stories is this Aristotelian one: you can't acquire the good for you without reasoning your way to it yourself.

[3] See Menn 2002, pp. 209–215, for the connection between the method of analysis and Plato's method of hypothesis from Meno 86e4–87b2.

Why can't the vicious person reason well?

Unlike Hume, Aristotle severs the question of whether reason is, from the question of whether it ought to be, a slave of the passions. In the *Topics*, he says, "the reasoning faculty does not always command, but sometimes also is under command, nor is that of desire and spirit always under command, but also on occasion assumes the command, whenever the man's soul is vicious" (129a15–18). Aristotle thinks that the vicious person's (and only the vicious person's) reason is enslaved, because he thinks that defects in his appetite and spirit make it impossible for the vicious person's reason to function well.

Why does Aristotle think that vice impedes deliberation? I present a two-premise argument for this conclusion:

(P1) Vicious people have, as their ends, things that are not good for them

Suppose one loves wealth greedily and indiscriminately. If such a person deliberates in the service of wealth, Aristotle says that he will get something bad:

> wisdom does not attend upon the unjust man. For the goods which he chooses and for which he commits injustice are what are absolutely good, not what are good for him. For wealth and office are good absolutely (haplōs), but for him perhaps they are not good; for by obtaining wealth and office he will do much evil to himself and his friends, for he will not be able to make a right use of office."
>
> (*Magna Moralia* II.3.7, 1199b9; cf NE I.2, 1094b15, NE V, 1129b5. EE VII.15, 1248b30)

The wealth that was his end was not in fact good for him, though it appeared to be.

The person who loves safety without having the virtue of courage deliberates her way to a kind of safety that is not good for herself; likewise, the one who loves honor without greatness of soul, the one who loves pleasure without moderation, the one who loves victory without gentleness. Aristotle acknowledges safety, honor, pleasure and victory as "good absolutely" but cautions people not to assume but rather to *pray* that "the things that are good absolutely (*haplōs*) may also be good for them" (NE V.1, 1129b2–8). The things that are good absolutely are good for a good person, in the way in which foods that are good absolutely are good for a healthy person, whereas punishment and medicine are good for a vicious or unhealthy person, respectively.

When the just person pursues pleasure, he is sure to end up with something good. His justice makes him such as to acknowledge that in a given case it can be *worse* for a person to have wealth or pleasure than not to have them: "pleasures are desirable, but not from *these* [i.e. disgraceful] sources, as wealth is desirable, but not as the reward of betrayal, and health, but not at the cost of eating anything and everything" (NE X.3: 1173b26–8). The just person won't seek those pleasures, and thus whenever he pursues pleasure, he pursues something good for himself. The goodness of ends such as wealth or pleasure (or: health, honor, victory, political office, friendship) can be undermined

both by how they will be used and the manner in which they are sought; the person who knows how to seek them is the one for whom they are good.

The virtues are the conditions in which the things that are good absolutely are good for us. They incline us to pursue wealth, health, victory and the rest under the right circumstances. That is why, when virtuous people deliberate about how to get wealth, or health, or victory, they are deliberating about how to get things that really are good for them, whereas vicious people are deliberating about how to get things that only appear good to them.[1] Suppose that we grant to Aristotle that the vicious person is in error regarding the end – that is he is wrong to want what he wants. Why does the fact that vicious people have the wrong end prevent them from deliberating well in the service of that end? Can't you be good at reasoning about how to get something bad? No, not when you have Aristotle's view of what deliberating is.

[1] Virtuous people have the "situational appreciation" (p. 237) described by Wiggins 1980: money strikes them as the thing to go for in those circumstances in which it really is.

(P2) Good deliberation preserves the status of your end as something good for yourself

We saw that success in derivation, in both the geometric and the ethical case, entails keeping hold of the distinctive (geometrical or ethical) status of the objects about which one is reasoning. In the geometrical case, this means that one must grasp the objects – the square and the circle and all the lines one adds – as geometrical objects rather than as markings on paper. One cannot, for instance, allow oneself to be distracted by the fact that it *looks* easy to draw a diagonal of a given circle with a ruler and end the analysis at A4. Because one has not yet found the center of the circle by any geometrical procedure, any drawing of what looks like a diameter capitalizes on physical properties of the drawn object that are not geometrical properties. If we lose hold of the status of the circle as a geometrical object, our "construction" will fail to be anything more than a drawing.

In order to *derive* the square, we must bring it about while preserving its distinctive normative status – which is to say, its status as a geometrical object. In the case of ethical reasoning, we must preserve the ethical status of the end, which is to say, its goodness for the agent. If one's end is wealth, one reasons in such a way as to preserve wealth as a *good* for oneself throughout the reasoning: for instance, one doesn't (usually) have any interest in acquiring wealth at the cost of one's life. If such a person fastened on a life-destroying means, that would render their practical reasoning defective.

C: The vicious person cannot reason well, because you can't preserve what is not there

The vicious person's thought is necessarily shot through with erroneous moves, because his end is something that is not, in fact, good for him. He cannot preserve the ethical status of wealth in his reasoning because, for him, it doesn't have any: wealth isn't good for him. But it is equally true to say that it doesn't have any ethical status because he doesn't know how to preserve it. His grasping, unjust attitude towards wealth prevents him from seeing the good of it, and this blindness explains both his misconceptions about the appropriate ways to acquire wealth and his misconceptions about what to do with wealth.

Deliberative excellence is a form of intellectual excellence that can only be manifested by people who grasp the good end in such a way as to be able to derive an actual good from it. This is what the ethical virtues allow a person to do. They are that condition of the passions which make sound reasoning in the service of one's end possible. In loving wealth without justice or generosity, the vicious person misses what is good about wealth. He mis-loves it. Like any fumbler or fool, his attempts to get and use wealth will therefore end up harming instead of helping him.

To say that the vicious person cannot deliberate well is not to say that the vicious person is completely bereft of intellectual resources he might deploy in acting. Aristotle acknowledges that the unjust man may end up with money and that he may even do so as a result of having "correctly" observed that money is attainable by theft. Nonetheless, Aristotle observes, (*NE* VI.9, 1142b17) "there is more than one kind of correctness." The sort of correctness of thought that tells the wicked man that he "can" get money by theft is mere Humean cleverness at causally linking ends and means. Aristotle calls this kind of cleverness "*deinotēs*" (*NE* VI.12, 1144a23–4). Merely clever thinking does not reflect a practically sound deliberative procedure: theft is not a rational way to get wealth, because it does not preserve the goodness of the end, wealth. The mistake is exposed when we observe that by such "correctness," the wicked person will "have gotten himself a great evil" (*NE* VI.9, 1142b20). The vicious person cannot distinguish good-preservative moves from erroneous ones. He cannot rationally discover the means to his ends.

It is, therefore, a mistake to understand the vicious person as making perfectly good calculations in the service of a defective goal. In Aristotle's theory of deliberation, there is something wrong with the calculations themselves. This is a point worth emphasizing, because Aristotle's claim that vicious people cannot reason well is not the much weaker claim that some reasoning "counts as" bad reasoning because it is done in the service of a bad end. Aristotle thinks the reasoning is bad on its own terms, having been made bad by the badness of its end. Compare: The iron which has been heated by a fire is hot because the fire was hot, but it now it has a hotness all its own.

Think back to the person who ends the analysis at A4 because she thinks she can just draw the diameter. She might take care to make sure that the line she is drawing is "correct" by making meticulous measurements with a ruler. But in pursuing this form of "correctness," she is not following any geometrical rules. Her failure to grasp the circle as geometrical translates into a failure to grasp any distinctively geometrical form of correctness.

If the passions are mal-educated, then when the person takes as his end things like health, wealth, safety, victory and honor, those ends will not in fact be good for him. An unjust desire for wealth or a cowardly desire for safety sends reason off on a wild goose chase, bent on "preserving" a good that is not there. Such reason cannot reason well; the best it can do is find a Humean path to bringing about wealth.

Hume thinks that the only practical use of reason is to plot a path to the satisfaction of passion's ends. But why does he think this condition counts as enslavement of reason? I do not "enslave" a rock when I use it as a door stop; I do not even enslave a chair if I use it as a table or as kindling. If something lacks ends of its own, I cannot do violence to it by using it for my ends. Contrast Aristotle's characterization of the vicious person, whose reason is alienated from its very capacity to function in accordance with its proper mode of operation. Reduced to seeking out causal connections to the object of desire, such a person's reason seems to have become a caricature, a shadow of its true self. It is intelligibly characterized as alienated from its own end of reasoning well. The shock value of Hume's claim that reason is "enslaved," that is, permanently divorced from its own ends, capitalizes on his audience's assumption that practical reason has ends of its own. But if Hume were speaking strictly and soberly, he would have to acknowledge that practical reason, as he understands it, *cannot* be enslaved.

Aristotelian deliberation is like geometrical reasoning: it has its own *sui generis* set of rational constraints that dictate what counts as a derivation in that domain. This is why Aristotle, unlike Hume, is in a position to assert that reason can be enslaved to the passions.

Deliberating immersively, not reflectively

Let us end by reflecting on the import of the difference between Aristotle's conception of deliberation and an evaluative one. Consider the phenomenon John Dewey called "reflective thought." Dewey, who popularized the phrase in his 1910 book *How We Think*, characterizes reflective thinking as "a conscious and voluntary effort to establish belief upon a firm basis of reasons" (p. 6). The reflective thinker resists her own immediate inclination to judge that some proposition is true or some action worth doing; instead she investigates whether the weight of the theoretical evidence or practical reasons supports the judgment she was inclined to make. She has a skeptical or self-critical attitude borne of a distinction between grounded and ungrounded judgment. In theoretical reasoning, reflection is an alternative to "jumping to conclusions," and in practical reasoning, reflection is an alternative to acting in a way that is not supported by the overall weight of the reasons.

The agent who engages in the evaluative deliberation familiar to us is someone who reflects on whether she ought (really) to do what she is inclined to do. Consider Christine Korsgaard's famous description:

> For our capacity to turn our attention on to our own mental activities is also a capacity to distance ourselves from them, and to call them into question... . I desire and I find myself with a powerful impulse to act. But I back up and bring that impulse into view and then I have a certain distance. Now the impulse doesn't dominate me and now I have a problem. Shall I act? Is this desire really a reason to act?
>
> (1996, p. 93)

Many who might prefer to think of deliberation in less introspective terms than Korsgaard would nonetheless agree with her that when we deliberate, we "back up" and ask ourselves whether the thing we were already inclined to do is *really* supported by the weight of the reasons. Deliberation, thus understood, interrupts action for a moment of potentially self-corrective reflection. It ensures that over and above simply acting, one acts in some especially well-grounded way: reflectively, or in a way that one can endorse, or in a way that satisfies some special procedural norm (i.e., so as to maximize expected utility). Evaluative deliberation is not required for acting *as such*; it is required only for acting *in that justified way*.

Reflective thinking is by its nature secondary to some process which has generated the options to be reflected upon. Dewey observes that reflective thinking

> begins in what may fairly enough be called a forked-road situation, a situation ... which proposes alternatives... . In the suspense of uncertainty, we metaphorically climb a tree; we try to find some standpoint from which we may survey additional facts and, getting a more commanding view of the situation, may decide how the facts stand related to one another.
>
> (p. 11)

When Korsgaard's agent "backs up," or when Dewey's "climbs a tree," they are responding to options that were already there. This is true whether, like Korsgaard's agent, we are independently inclined to do some one thing or whether, like Dewey's, we simply see the road forking and are unsure which way is best.

Evaluative deliberation is only possible for someone who could have done something without deliberating. She could have simply followed inclination, or, in the forking case, she could have made an arbitrary choice. (Arbitrary choice may in any case be the only solution in a Buridan's ass case where one's options are virtually identical. Likewise, one might have to pick arbitrarily if one runs out of time to deliberate or if the difference between the options does not warrant a deliberative time-investment.)[1]

Aristotelian deliberators are in a different predicament. They deliberate in search of something (anything) to do in the service of their end. As such, their thinking is not reflective but rather, as I will call it, *immersive*. A familiar example of immersive (though not deliberative) thought is provided by the detective who is pondering a puzzling crime just as her eye falls on the crucial detail. Suddenly, she finds herself constructing a complete account of what happened, fitting together various details that had previously struck her as disconnected or irrelevant. Consider, in this connection, the activity of trying to recollect something. When one seeks to remember a detail of, for example, a story or movie or past experience, one brings to bear a gradually more detailed reminiscence of elements relating to the missing one until the relevant part 'clicks.'

In immersive thought, one cannot separate the question of whether some answer is the *right* answer from the question of whether it is an answer at all. For this reason, the thinker's positive assessment of an answer must be internal to the very process by which she generates the answer: she senses that the relevant move seems right or familiar or that the elements fit together or that she has got it. She experiences her thinking as going well but not because she can assess it by a separable standard of success:[2] instead, she simply feels that she is connecting or associating things that

[1] For a discussion of the rationality of arbitrary choice, see Ullmann-Margalit and Morgenbesser 1977.

[2] Thus, Aristotle describes the ultimate moment of deliberation as consisting in a moment of quasi-

belong together. In Stephen Menn's discussion of analysis, he returns again and again to a certain way of describing moments like the identification of the chord at A5: "And then at some point something clicks" (p. 195, 198, 199, 217). Aristotle sees this noetic moment of insight as a common point between analysis and deliberation: they end in one just "seeing" the answer (VI.8). By contrast, in reflective thinking, one can take what has already been understood as a candidate answer and ask whether it meets some independently graspable standard for being the best answer. There's no "clicking" because those are two things, not one.

If you have available to you an option that would achieve your end, then you have no problem that Aristotelian deliberation could solve. The person who knows how to construct a square in a circle will, upon being presented with the problem, go ahead and produce the synthesis starting from, for example, S1. Such a person *cannot* use the Aristotelian analytic method to derive the starting point, because she already knows the starting point. She may go through the relevant motions, drawing the sequence A1–5, but this "analysis" must be a sham: you can at best *pretend* to search for what you already have.

The Aristotelian deliberator could not have done anything without deliberating. The evaluative deliberator may not have acted *as well* if she had not deliberated, but she could have acted. She was deliberating between things that she could (already, pre-deliberatively) have done. She had options.

The evaluative approach to deliberation understands the process – whether it be in the form of comparing or testing – to be a kind of *checking*. When we evaluate some proposed option, we are taking a certain deliberative initiative. Dewey says reflection is a product of a "conscious and voluntary effort." The action we select for evaluation is not, of course, selected at random. It is something we see some reason to do. The question we ask is: Do we have the *most* reason to do it? Is it the *best* thing we can do? Is it better than this or that alternative action I also see some reason to do? Does it pass every relevant moral test? Evaluative deliberation seems to be born from doubts as to whether the action that it occurs to a person to do might be the wrong thing to do. Thus we are called upon to reflect, to "weigh reasons," to form "all things considered" judgments and test proposed actions to see whether they could become universal laws of nature. Carried to its extreme, such deliberative work is sometimes taken to have the potential to re-orient a person, so that she reasons her way out of her current ends and values and into ones of a radically different kind.

In Aristotle's picture, there is no neutral standpoint from which a person might, rising above the fray of his own self, rationally call his impulses, values and projects into question. Aristotelian deliberation is unashamedly pragmatic,[3] tasked with finding an action when none immediately presents itself. Aristotle's account of deliberation

perceptual intellectual insight by which the person grasps, for example, that the chord is the right line (*NE* VI.9, 1142a25–30).

[3] It is important to keep in mind, however, that excellence at deliberation does not exhaust intellectual virtue – it does not even exhaust practical intellectual virtue. For there is, in addition, the virtue

thus throws our own concern with practical grounding into relief. By reading him, we come to appreciate how profoundly our own approach to practical thought is shaped by the felt need to respond to anxieties about practical justification – anxieties that Aristotle reveals are not written into the very concept of reason made practical.

of comprehension (*sunesis*), discussed in VI.10, and forgiveness (*suggnome*), discussed in VI.11. For an account of the relation of those virtues to deliberation, see Segvic 2009, pp. 160–162.

Works cited

Barnes, Jonathan, ed. (1984). *The Complete Works of Aristotle*. Princeton: Princeton University Press.

Broadie, Sarah (1991). *Ethics with Aristotle*. New York: Oxford University Press.

Cooper, John (1986). *Reason and Human Good in Aristotle*. Indianapolis: Hackett Publishing.

Davidson, Donald (1980). "How Is Weakness of the Will Possible?" In his *Actions and Events*. Oxford: Clarendon Press.

Dewey, John. (1910). *How We Think*. Boston: Heath et. Co.

Hume, David. (1739). *A Treatise of Human Nature*, 1896 reprint ed. L. A. Selby-Bigge, M.A. Oxford: Clarendon Press.

Korsgaard, Christine (1996). *The Sources of Normativity*. Cambridge: Cambridge University Press.

Kosman, Aryeh (1980). "Being Properly Affected: Virtues and Feelings in Aristotle's Ethics," in *Essays on Aristotle's Ethics*, ed. A. Rorty. Berkeley: University of California Press.

McDowell, John (1979). "Virtue and Reason," *The Monist* 62 (3): 331–350.

Menn, Stephen (2002). "Plato and the Method of Analysis," *Phronesis* 47.

Moss, Jessica (2011). "Virtue Makes the Goal Right: Virtue and *Phronēsis* in Aristotle's Ethics," *Phronesis* 56.

Nielsen, Karen (2011). "Deliberation as Inquiry: Aristotle's Alternative to the Presumption of Open Alternatives," *Philosophical Review* 120.

Normore, Calvin (1998). Picking and Choosing: Anselm and Ockham on Choice," *Vivarium* 36.

Nussbaum, Martha (2001). The Fragility of Goodness*:* Luck and Ethics in Greek Tragedy and Philosophy*. Cambridge: Cambridge University Press.*

Price, Anthony (2011). "Aristotle on the Ends of Deliberation," in *Moral Psychology and Human Action in Aristotle*, eds. Michael Pakaluk and Giles Pearson. Oxford: Oxford University Press.

Rowe, Christopher, trans. (2002). *Aristotle, Nicomachean Ethics*, ed. Sarah Broadie. Oxford: Oxford University Press.

Segvic, Heda (2009). "Deliberation and Choice in Aristotle," in her *From Protagoras to Aristotle*. Princeton: Princeton University Press.

Sorabji, Richard (1980). "Aristotle on the Role of Intellect in Virtue," in *Essays in Aristotle's Ethics*, ed. A. O. Rorty. Berkeley: University of California Press.

Ullmann-Margalit, Edna, and Sidney Morgenbesser (1977). "Picking and Choosing," *Social Research* 44.

Wiggins, David (1980). "Deliberation and Practical Reason," *Essays in Aristotle's Ethics*, ed. A. O. Rorty, 221–240. Berkeley: University of California Press.

9. Hume's robust theory of practical reason[1]

[1] I am extremely grateful to Don Garrett for fun and helpful conversations about Hume on practical reason and for careful comments on an earlier draft of this chapter. I am grateful too for very valuable and detailed feedback from Ruth Chang and Karl Schafer, as well as for help from audiences at The Ohio State University, NYU/Abu Dhabi, Kings College London, the University of St. Andrews, and the Rocky Mountain Ethics Conference at the University of Colorado, Boulder. References to Hume's work in the body of this chapter use "*Treatise*" to refer to *A Treatise of Human Nature* (1739–1740), "*Enquiry,*" to refer to *An Enquiry Concerning the Principles of Morals* (1751), and "*Sceptic*" to refer to "The Sceptic" (1742).

Introduction

Hume never uses the phrase "practical reason." This is no surprise, given his commitment to revealing the unfounded pretensions of those who appealed to "reason" as an all-purpose safe haven for their preferred views of theology, science, or morality.

Yet Hume clearly has a great deal to say about practical reason. In light of what he says, Hume is regularly read as either an outright skeptic about practical reason or as an advocate of unadorned instrumentalism. According to the skeptical reading, Hume rejects the idea that reason could be practical at all. According to the instrumental reading, he embraces reason as practical yet sees its role as being entirely a matter of figuring out efficient ways to satisfy one's desires or achieve one's ends.[1] The instrumentalist interpretation has become so widespread that instrumentalism is often labeled 'Humeanism' (though, in a nod to the plausibility of the skeptical reading, people often say that it is unclear whether Hume is a "Humean").[2]

Not surprisingly, support for both interpretations is easy to find. So, for instance, when it comes to the case for the skeptical interpretation, people point to Hume's claim that "reason is perfectly inert, and can never either prevent or produce any action or affection" (*Treatise* 3.1.1.8, SBN 457–458), which seems an unambiguous assertion of the skeptical view. Meanwhile, those interpreting Hume as an instrumentalist point to Hume's assertion that "Where a passion is neither founded on false suppositions, nor chooses means insufficient for the end, the understanding can neither justify nor condemn it" (*Treatise* 2.3.3.6, SBN 415–6), which seems almost as clearly to highlight explicitly just what an instrumentalist would identify as practical reason's role – to ensure that our desires are informed and that we choose effective means to their satisfaction.

The details of the arguments offered for these interpretations vary, with some appealing to Hume's theory of action, others to his rejection of reason as the source of moral distinctions, and still others to his theory of meaning. In each case, I will be arguing, these interpretations end up missing the subtle, detailed, and plausible theory of practical reason that Hume develops. My aim here is to bring out that robust theory.

[1] Defenses of the skeptical interpretation include Christine Korsgaard (1986), Jean Hampton (1995), and Elijah Millgram (1995). Elizabeth Radcliffe (1997), in contrast, defends an instrumentalist interpretation. See Kieren Setiya (2004) and Karl Schafer (2015) for important exceptions to the rule that people see Hume as either a skeptic or an instrumentalist.

[2] For example, see Michael Smith (1987), David Lewis (1988), and Donald Hubin (1999).

Elements of a theory of practical reason

Any non-skeptical theory of practical reason needs to do at least three things:

First, it must give an account of the *activity* of practical deliberation – of what is involved in facing a situation, canvassing options, weighing considerations that one takes to count in favor or against various options, and then successfully acting in light of, and on the basis of, what one sees the balance of the considerations as supporting.

Second, it must identify the *capacities* one must have in order in order to engage in that activity, specifying what faculties, capacities, or abilities are required in order to be able to engage in practical deliberation.

Third, it must articulate and defend the *standards* to which those engaging in the activity are properly subject, in light of which they count as practically rational (or not), the considerations count as reasons (or not), and the conclusion they reach as justified (or not).

Hume, I will argue, does all three. Before turning to that argument, though, it is important to note that those who read Hume as a skeptic themselves need to say enough about one or another of these three aspects of practical reason in order to make the case that Hume was a skeptic about (that aspect of) practical reason.

Some argue that Hume's skepticism lies in his leaving no room, in fact, for people to think of various considerations as counting in favor (or against) various options and so no room for people to act in one way or another on the basis of the conclusions such thoughts would support. On this view, Hume holds that while we have all sorts of beliefs about various things that might – and often do – result in our behaving in various ways, we neither do nor can have beliefs to the effect that these considerations count in favor or against different courses of actions, so we can never act as we do for what we take to be reasons.

Others hold that Hume's skepticism lies in rejecting the idea that reason could possibly cause actions, which they take to be a *sine qua non* of *practical* reason. On this view, Hume's account of our capacities, and specifically his understanding of reason's powers, commit him to rejecting the idea that reason (and so deliberation about one's options) could cause, or stand in the way of, action. This ensures, they argue, that reason, as Hume understands it, cannot possibly be practical.

Still others maintain that Hume was a skeptic about there being any standards in light of which our deliberations, choices, and actions might appropriately be judged as rational (or not) and no standard in light of which some considerations and not

others are properly seen as reasons for or against doing various things. How else, they ask, are we to explain Hume's assertion that "'Tis not contrary to reason to prefer the destruction of the whole world to the scratching of my finger" (*Treatise* 2.3.3.6, SBN 415–6)? On this view, whether or not we engage in practical deliberation, and whether or not we have the capacities required for deliberation to be effective, there are no standards, on Hume's view, for our preferences, choices, deliberations, or actions: "Actions may be laudable or blameable," Hume notes, but "they cannot be reasonable or unreasonable" (*Treatise* 3.1.1.10, SBN 458).

Those who interpret Hume as an instrumentalist join me in rejecting the skeptical alternative. Together we credit Hume with an account of the activity of practical deliberation and the capacities that activity requires and also credit him with an account of the standards in light of which the activity, and the resulting actions, might appropriately be judged rational (or not). Where we part company is in our understandings of the accounts Hume offers. I will be arguing, in what follows, that in all crucial respects, Hume is not an instrumentalist. Hume does of course allow that the recognition that some course of action would satisfy one's desires, or would be a means to achieving one's ends, often gives one reason to perform the action and serves to make the action (if performed for that reason) practically rational – though, as I mentioned previously, he does not use that language. Yet Hume rejects both the idea that all reasons for action are found in the action's role in satisfying desires or achieving ends, and the idea that the only normative standards are instrumental. Like Kant, he recognizes a central role for instrumental reasoning while situating it within an account of when and why various desires and ends do – and when they do not – underwrite reasons for acting in ways that will satisfy or achieve them.

To make the case, I will take the three elements of a theory of practical reason – an account (i) of the *activity* of practical deliberation, (ii) of the *capacities* the activity requires, and (iii) of the *standards* that apply to those engaging in the activity – in reverse order.

Hume on the *standards* for practical deliberation

The first thing to say here is that Hume is unmistakably a skeptic about *rationalist* defenses of various standards for practical deliberation. In the *Treatise*, he argues at length that moral distinctions – between virtue and vice, right and wrong – are not "discern'd merely by ideas, and by their juxta-position and comparison" and so (he concludes) not derived by reason alone.

These arguments play out against the view that reason's domain is restricted to canvassing and comparing ideas, to discovering relations that might stand between those ideas and their object, and to using demonstrative and probable reasoning in order to judge various ideas true or false.[1] When only reason is involved, the judgments of truth and falsehood it supports depend entirely on demonstrative reasoning and what it can establish without appeal to experience. To show that moral distinctions are not based solely on reason, Hume moves through what he takes to be the only strategies that demonstrative and probable reasoning would make available, arguing that each fails to generate the sort of substantive standards that are evidently involved in distinguishing virtue from vice and right from wrong.

In his critique of rationalism and his subsequent development of an alternative account of the standards of practical deliberation, Hume does not consider views of the sort Kant defends, according to which it is an analytic truth that "Whoever wills the end, also wills (insofar as reason has decisive influence on his actions) the means that are indispensably necessary to it that are in his control."[2] If such a claim is analytic, it is just the sort of thing reason alone is, on Hume's view, able to discover.

So, it is worth noting that Hume could, without any problem, grant the truth of Kant's claim. For a certain understanding of 'reason', and of what it is for reason (so

[1] In doing this, Hume is following Locke, who wrote that reason is "the discovery of the certainty or probability of such propositions or truths which the mind arrives at by deduction made from such ideas, which it has got by the use of its natural faculties; viz. by sensation or reflection" (Locke 1689, iv 18.2). He is not introducing some unfamiliar or unmotivated constraint on reason, though he notes that "reason" is often used loosely, to the detriment of understanding what it might or might not establish. In the *Enquiry*, Hume marks the same two modes of reasoning but refers to the second, the determining of whether particular ideas are true or false, as moral reasoning: "All reasonings may be divided into two kinds, namely demonstrative reasoning, or that concerning relations of ideas, and moral reasoning, or that concerning matter of fact and existence." [*Enquiry.* 4.18, SBN 35]

[2] Kant explains the analyticity this way: "As far as volition is concerned, this proposition is analytic; for in the volition of an object, as my effect, is already thought my causality as an acting cause, i.e.,

understood) "to have a decisive influence" on one's actions, it may well be "discern'd merely by ideas, and by their juxta-position and comparison" that, when reason has decisive influence, in willing an end, one wills the "means that are indispensably necessary to it that are under his control."

At the same time, Hume could well – and presumably would – also argue that the standing of reason, so understood, as *setting a standard* for action is not settled by the analytic truth. The analytic truth is perfectly compatible, as Kant notes, with reason not actually having a decisive influence on one's will. Granting the truth leaves open the question of whether reason (so understood) should have that influence. Hume doesn't, and needn't, deny that it should. What Hume is committed to is holding that *this* question is not answered by the truth on offer nor by it being analytic and that whether the truth serves to set a standard cannot be settled by reason alone.[3]

Analytic truths establish what must be true. But in this case, the truth that is supposed to be settled – about the implications of reason, understood in a certain way, having a decisive influence on one's will – is silent on whether reason *should* have a decisive influence and so on whether our will ought to be such as always, in willing an end, to will the means that are "indispensably necessary and within our power." In recognizing the analytic truth, there is no commitment one way or the other as to what standards, if any, are appropriate – unless, of course, one already valorizes the sort of rationality on offer. But in that case, what is called for is an argument that rationality of the kind in question is a virtue or that it sets a standard for our will. The proffered analytic truth does not provide that.

Hume's skepticism about *rationalist accounts* of the standards of virtue is, importantly, not a skepticism about there being true judgments of virtue and vice, nor, more generally, about there being genuine standards for practical deliberation. Indeed, Hume begins the *Enquiry* writing that

> Those who have denied the reality of moral distinctions, may be ranked among the disingenuous disputants; nor is it conceivable, that any human creature could ever seriously believe, that all characters and actions were alike entitled to the affection and regard of every one.
>
> (*Enquiry* 1.2, SBN 169–70)

---

the use of means" (Kant (1785), [Ak 4:417]). This would explain why, in willing the end, one is, simply in virtue of that, willing certain means to its achievement. But it is not clear that this is the claim Kant needs. He acknowledges that one may maintain an end and in fact not pursue what one recognizes as the necessary means to its achievement. This is a classic case of weakness of will, and he wants to account for it by saying that in these cases, reason does not have a decisive influence on one's will, which requires the analytic truth he is after to be one between reason having a decisive influence on one's will and one willing the (recognized) necessary means to one's end. Such a truth would presumably turn on the nature of the idea of reason that is in play.

[3] There is an understanding of Kant's claim that leaves no room for one to will an end without at the same time willing the indispensably necessary means to its achievement. It follows Kant's official explanation of the analyticity – willing the end as an end is willing the means – but it then abandons

Rather, Hume's skepticism is about reason's capacity to discover – alone and unaided by experience – the standard(s) in light of which such judgments would be true. Hume's conclusion is simply that "since vice and virtue are not discoverable merely by reason, or the comparison of ideas, it must be by means of some impression or sentiment they occasion, that we are able to mark the difference betwixt them" (*Treatise* 3.1.2.1, SBN 470).[4] In other words, experience must be added into the mix in order to account for the standards that distinguish virtue from vice. Hume ends up developing an account of the relevant standards that depends crucially on sympathy and the sentiments of approbation and disapprobation that sympathy makes available.

Hume frames his discussion of normative standards in terms of virtues and vices, not reasons and rationality. And much of his discussion focuses on what are commonly recognized as *moral* virtues and vices. So, it would be easy to think that Hume's account of the standard of virtue and vice is restricted to what we now think of as morality, without implications for practical rationality and the nature of reasons more broadly. But that would be a mistake. Hume's discussion comes with an explicit commitment to the category of virtues and vices being much broader than just the moral and as including traits that are commonly thought to fall within practical rationality but not morality. In particular, he mentions "prudence, temperance, frugality, industry, assiduity, enterprize, dexterity" as well as "*perseverance, patience, activity, vigilance, application, constancy*" and "other virtues of that kind" (*Treatise* 3.3.4.7, SBN 610–11). As Hume develops his account of the relevant standards, he is concerned with identifying, articulating, and defending them as standards of, as he put it, "personal merit," without regard to whether they would normally be thought of as moral.

Indeed, in "Of Some Verbal Disputes" (an Appendix to the *Enquiry*), Hume makes a special point of holding that there is a single general story of the standards in

---

the idea that we should will the means but may not, which is crucial to the idea that we are talking about a normative standard we should, but at least in principle might not, meet. Kant's framing of the claim itself makes clear that he recognizes a failure is possible: reason, as he acknowledges, might not have a "decisive influence." In fact, Kant notes that imperfectly rational beings, which includes all people, are such that what is objectively necessary is for them subjectively contingent.

[4] Hume goes on to finish the thought, writing "Morality, therefore, is more properly felt than judg'd of." People have often taken this as grounds for thinking Hume rejected the idea that we make moral judgments, which might be true or false, and instead embraced emotivism or some other form of non-cognitivism. But that mistakes his point here, which is that our moral judgments require the input of experience (in the form of either an impression or a sentiment, both of which are felt, not judged). At the same time, it flies in the face of his careful development of a theory of the standard virtue in light of which our moral judgments are to be evaluated as true or false. Moreover, Hume makes clear that once the standard is in place, we can both make true moral judgments, without the corresponding feeling of approval, and have feelings of approval for what is not, in fact, virtuous. In these respects, our judgments of virtue and vice are analogous to our judgments of color. Such judgements are possible for us only because of certain kinds of experiences people are able to have, and they need to be understood in terms of those experiences, but the truth of the judgments we might make concerning them are independent of the particular experiences we might have, even though we often rely on our experiences (of approval and of visual experience) in making those judgments.

light of which people should be judged meritorious, even if different standards come into play (within that story) regarding different aspects of peoples' character, motives, and actions. And he emphasizes again that these include standards for what are now commonly thought a matter of practical rationality, not morality.

According to Hume's overarching story, traits of character, motives, and actions count as meritorious (i.e. as virtuous) *if, when, and because* they would secure approval from an appropriately "common" or "general" point of view.[5] In the *Treatise*, he goes into great detail about what that point of view is like, about why it is needed, and about why it sets the appropriate normative standard for our judgments of virtue and vice. On Hume's account, taking up the appropriate general point of view requires (i) knowing the relevant facts about who is likely to be affected (and how) by the traits in question, (ii) being engaged via sympathy with the welfare of those who are affected, and (iii) responding with approval (or disapproval) without regard to one's own interests. In short, Hume holds that the standard for our judgments is set by the reactions of those who are appropriately informed, genuinely concerned, and yet suitably impartial. Their reactions set the standard for distinguishing between what happens to secure approval and what is actually approvable.[6]

In defending his account, Hume requires that the standard it offers line up reasonably well with the distinctions we actually draw between virtue and vice, since otherwise it would not plausibly count as the standard for what we are actually thinking about. Yet, at the same time, he insists that the correct standard for those distinctions must be one that we can justifiably endorse as appropriate – it must itself be approvable and not merely be something that captures what we happen to approve of. As Hume puts it at the end of the *Treatise*,

> not only virtue must be approv'd of, but also the sense of virtue: And not only that sense, but also the principles, from whence it is deriv'd. So that nothing is presented on any side, but what is laudable and good.
>
> (*Treatise* 3.3.6.3, SBN 619)[7]

[5] (*Treatise* 3.3.1.30, SBN 590–1 and *Enquiry* 9.6, SBN 272–3). I have been lumping character, motives, and actions together in this discussion so as to leave things general. It is worth noting, though, that on Hume's view, the relevant aspect of a person's character is constituted by her motivations, which determine both the quality of her character and the quality of her actions. He thinks well-motivated attempts at various actions are fully virtuous and that actions that are ill motivated, no matter what their effects, are not approvable, however salutary their impact. In this way, Hume is much like Kant in finding the full value of how a person carries herself in her motives (including, of course, her appropriate concern for others) and not in the actual effects of what those motives happen to cause.

[6] Similar accounts, albeit with important differences, have been offered by Adam Smith (1790), Roderick Firth (1951), and, most recently, Michael Smith (1994). What they share is the idea that the appropriate standards for our normative judgments of morality or rationality are set by the reactions of those who are suitably situated, even as they differ in important ways about which reactions matter and what it takes for someone to count as appropriately situated.

[7] In insisting on this point, Hume is taking issue directly with Hutcheson, who claims that the

We can leave to one side most of the details of Hume's account of what he takes to be the appropriate point of view, which we might call the "General Point of View," given the aims of this chapter.[8] But it is crucial to register three things.

The first is that Hume embraces the General Point of View as setting the standards for the decisions and actions that are usually seen as falling within the ken not just of morality but more generally of practical rationality. Hume was not a skeptic about there being such standards. Moreover, he devotes a lot of time and attention to articulating and defending how and why the General Point of View sets the standards it does.[9]

The second is that, on Hume's account, the standards set by the General Point of View are not purely instrumental, even when considering prudence. The standards do, in many cases, treat motives and actions that promote the satisfaction of the agent's desires as meritorious. This is because those who take up the General Point of View and so focus, sympathetically and impartially, on what will advance the welfare of the agent in question will often approve of that agent acting in ways that will satisfy his or her desires. But not always, and for two reasons. The first is that sometimes acting to satisfy one's desires will predictably not make one better off, and when that is the case, so acting will not secure approval from those who take up the General Point of View.[10] The second is that in taking up the General Point of View, one will, from an informed concern for the agent's welfare, approve of certain motives and actions regardless of whether the agent has a relevant desire. So, while Hume's standard for prudence takes an agent's desires seriously, it treats them as neither sufficient nor necessary for an agent to have prudential reason to act in certain ways. Moreover, as attention shifts from prudence to other traits of character, what will secure approval from the General Point of View will not be settled by the agent's desires, even as, often, the agent's desires, as well as the desires of others, play a central role in what reasons we have (when and because they help to determine what would be approved of from the General Point of View).

The third is that once the General Point of View, and the standards it makes possible, are in place, Hume is in a position to mark the crucial normative distinction between what *happens to secure* our approval and what actually *merits* that approval.

---

sense of virtue neither admits of, nor requires, a defense as itself justified. See Hutcheson (1742), sec. 1. Adam Smith follows Hume and offers a sustained defense of the position they share. See Smith (1790), p. 323. I explore this requirement in Sayre-McCord (2013).

[8] Hume starts with the observation that common or general points of view are a central part of making judgments of virtue and then works to identify the specific general point of view that informs our judgments. For some details about Hume's view of the character of that specific general point of view, see Sayre-McCord (1994, 1996, 2013).

[9] Something similar can be said about Hume's defense of certain standards of causal reasoning, which he clearly sees as standards by which we should be judged and to which we should conform. See by Deborah Boyle (2012), for an interesting discussion of the relation of these rules to the virtue of wisdom.

[10] More accurately, if such actions might secure approval from the General Point of View, it will not be because of the (expected) satisfaction of those desires but because of some other benefit the actions generally promote.

This provides critical purchase on actual practice, allowing Hume to make sense of, say, the "monkish virtues" not being virtues at all and the various practices we might endorse being open to serious criticism and in need of substantial revision. While Hume is committed to the role of experience (including those experiences that result from our capacity for sympathy and feelings of approval) in determining the appropriate standards, he does not hold that those standards are a function of what people actually feel or approve of – nor are those standards simply whatever people might happen to think they are. Hume is attempting to capture what we are thinking and doing in making judgments of virtue and vice, and to that degree, his aim is descriptive, but the judgments he is giving an account of are not descriptions of what social practices and beliefs happen to be but of what they should be – they are not, as Hume might put it, judgments of what *is* but judgments of what *ought* to be.[11]

Still, just to fill in the details a little bit... . When it comes to benevolence, for instance, Hume's view is that its standing as a virtue depends on our vulnerability, on our potential need for the help of others, and on other's ability actually to help. With that vulnerability in mind, were we to take up the General Point of View, focus on how people can benefit from others trying to help, and sympathize with those who might thereby be benefitted while leaving aside our own interests, we would approve of the efforts to aid that are involved in acting benevolently. Yet were things different in crucial respects, for instance, if we were never in need of help, or if people could not help despite our need (and perhaps would make things worse if they tried), then benevolence, on Hume's view, would not be a virtue – precisely because, under such conditions, it would not receive the requisite approval from the General Point of View.[12]

[11] Hume is, of course, famous for arguing that in our reasoning, we need to register and explain the jump from claims about what is to what ought to be: "as this *ought*, or *ought not*, expresses some new relation or affirmation, 'tis necessary that it shou'd be observ'd and explain'd; and at the same time that a reason should be given, for what seems altogether inconceivable, how this new relation can be a deduction from others, which are entirely different from it" (*Treatise* 3.1.1.27, SBN 469–70). Hume's explanation is that the transition from the various claims about what is to conclusions about what ought to be are forged by the standards that would be endorsed – that is, approved of – from the General Point of View. Of course, he holds as well that an inference from what would be approved of to what ought to be calls for explanation, and he thinks it is an important virtue of his account that the explanation is found in the General Point of View also approving of itself setting the standard for our standards. If it did not, he holds, it would have to be rejected.

[12] From the General Point of View, one considers the trait in question (in this case, benevolence), with an eye to its usual effects (on the possessor and others) under standard conditions, where standard conditions are determined by the practical problems faced by the people whose characters are being considered. (There is a neat and complicated story behind the General Point of View having this focus, which I leave to one side here.) In having all that in view, and leaving aside one's own interests in favor of being influence solely by sympathy, when one considers the effects of the character, one ends up feeling either approval or disapproval directed at the people with the character in question in light of that character's effects on the weal or woe of those being considered. So in considering benevolence's effects on people like us, from the General Point of View, we are moved to approve in light of the benefits benevolence usually has. But if we were in a world where, in fact, the road to hell is paved with attempts to help others, and we knew this, attempting to help others would not secure the requisite approval.

Similarly, when it comes to prudence, its standing as a virtue depends on the prospect people have of benefitting from a concern for their own welfare. To the extent people can and do benefit from such a concern, were we to take up the General Point of View, focus on how those who pursue their own interest fare, and sympathize with them while leaving aside our own interests, we would approve of their efforts to advance their welfare. Yet were things different in crucial respects, for instance, if a concern for one's own welfare consistently made one worse off, then such a concern, on Hume's view, would not be a virtue – precisely because, under such conditions, it would not receive the requisite approval from the General Point of View.

Just as the virtue of benevolence requires a respect for the overall welfare of those one is trying to help, a respect that may properly restrain efforts to help, so too the virtue of prudence requires a respect for the overall welfare of those trying to advance their own interests, a respect that may properly restrain those very efforts.[13] The virtuousness of the efforts, as well as the limits on them, are, according to Hume, explained by an appropriate appreciation of when and why the efforts would secure approval from the General Point of View.[14]

[13] Benevolence and prudence are not only, according to Hume, conditionally virtuous, when virtuous, they have their limits. A concern with helping others can quickly lead to meddling, paternalism, or dependency; a concern with promoting one's own welfare can stand in the way of other concerns that enrich one's life and repel others on whose affection one's welfare turns. Here too, Hume thinks, the General Point of View explains the limits.

[14] Most of Hume's detailed discussion of the standards of morality and prudence focuses on character traits and is tied to his view that the virtue or viciousness of an action depends on why a person performed it. But his general account of normative standards, and the role the General Point of View plays in that account, is fully compatible with his recognizing (as he does) that there is something good about a just action regardless of why a person performs it and something good about one meeting or conforming to other kinds of standards (of action or reasoning) independently of why one might be doing so. This leaves room for exploring whether, for instance, there are standards of practical reason that do not focus on character traits that might secure approval from the General Point of View. A standard that requires that one pursue what one recognizes as necessary to the achievement of one's ends or abandon the end might well fall in this category. As it happens, although Hume does offer standards for theoretical reasoning in Book I of the *Treatise*, he does not do the same with regard to practical reasoning.

Hume on the *capacities* required for the activity of practical deliberation

Hume highlights two capacities in particular as crucial to practical deliberation: reason and sentiment. Both, in his account, are required in order for someone to engage in the activity of practical deliberation.

In discussing reason, and its practical role, Hume relies on what he calls a "strict and philosophical sense" of reason (*Treatise* 3.1.1.12, SBN 459–60). So conceived (and here Hume is pretty much following Locke), reason's role is limited to determining truth and falsity on the basis of "demonstrative and probable reasonings" (*Treatise* 2.2.7.n, SBN 371). On this view, to be conformable to reason is to be true; to be contrary to reason is to be false. Two dramatic conclusions follow directly from this view of reason, as Hume points out.

The first is that

> 'Tis not contrary to reason to prefer the destruction of the whole world to the scratching of my finger. 'Tis not contrary to reason for me to chuse my total ruin, to prevent the least uneasiness of an *Indian* or person wholly unknown to me. 'Tis as little contrary to reason to prefer even my own acknowledg'd lesser good to my greater, and have a more ardent affection for the former than the latter.
>
> (*Treatise* 2.3.3.6, SBN 415–6)

This is because preferences and choices cannot themselves be true or false.[1] Of course denying that these preferences and choices are contrary to reason – that is, are not themselves false – is perfectly compatible with holding that we should not have

[1] The way Hume puts this point is that each is an "original existence" and "contains not any representative quality, which renders it a copy of any other existence or modification... . 'Tis impossible, therefore, that this passion can be oppos'd by, or be contradictory to truth and reason" (*Treatise* 2.3.3.5, SBN 415). Hume repeats the point in Book III: "Now 'tis evident our passions, volitions, and actions, are not susceptible of any such agreement or disagreement; being original facts and realities, compleat in themselves, and implying no reference to other passions, volitions, and actions. 'Tis impossible, therefore, they can be pronounced either true or false, and be either contrary or conformable to reason" (*Treatise* 3.1.1.9, SBN 458).

these preferences and should not make these choices. Moreover, the denial is perfectly compatible with thinking that there are true judgments, discoverable by reason, to the effect that these preferences and choices are vicious, wrong, or bad or such that we have reasons (perhaps decisive reasons) not to have or make them. Hume's point is that the truth of such judgments cannot be explained by the preferences or choices themselves being contrary to reason, since they are not the sort of thing that can be false (or true).[2]

The second dramatic conclusion is that "reason can never immediately prevent or produce any action *by contradicting or approving of it"* (*Treatise* 3.1.1.10, SBN 458, ital. added).

This is because actions, no less than preferences and choices, cannot be true or false and so can neither be contradicted, nor approved of, by reason. Consequently, reason can neither cause actions, nor prevent them, "by contradicting or approving" of them.[3]

Hume goes on from there to contrast reason, in this regard, with morality, which *can* cause actions by contradicting or approving of them. "The merit and demerit of actions frequently contradict, and sometimes controul our natural propensities. But reason has no such influence" (*Treatise* 3.1.1.10, SBN 458). The difference between reason and morality that matters here is in their respective means of contradicting and approving – truth and falsity in the first case, merit and demerit in the second –

[2] Hume is here arguing directly against the eighteenth-century British moralist William Wollaston, who maintains that the immorality of actions was to be explained by those actions making false assertions. He claims that there are many acts, including those "such as constitute the character of a man's conduct in life, which have *in nature*, and would be taken by any indifferent judge *to have a signification*, and *to imply some proposition*, as plainly to be understood as if it was declared in words: and therefore if what such acts declare to be, is not, they must *contradict truth*, as much as any false proposition or assertion can." Their immorality, he argues, is found in their falsity. (See Wollaston (1722, Section 1, III, p. 7).) Against the background of this argument, Hume spends some time exploring the idea that the viciousness of such preferences and choices can be traced either to the falsity of the beliefs that cause them or to the beliefs they cause, finding them all wanting. In each case, the problem is that the falsity of the candidate beliefs is neither necessary nor sufficient for the viciousness of the preferences and choices with which they might be connected as cause or effect.

[3] Alluding directly to this argument, Hume misleadingly characterizes the key conclusion in these terms: "reason is perfectly inert, and can never either prevent or produce any action or affection," leaving off the clause "by contradicting or approving of it" (*Treatise* 3.1.1.8, SBN 457–8). But without that clause, Hume is in no position to hold that reason can never either prevent or produce any action or affection. Indeed, experience alone seems to provide ample evidence that reason, whether we are talking about reasoning or about the products of reasoning, can and often does have effects – pleasures, headaches, conclusions – that might lead to actions or affections. Hume of all people would not hold *a priori* that they do not, having argued at length that "there are no objects which by the mere survey, without consulting experience, we can determine to be the causes of any other; and no objects, which we can certainly determine in the same manner not to be the causes. Any thing may produce any thing" (*Treatise* 1.3.15.1, SBN 173). What he can and does hold *a priori* is that reason cannot have this effect *by contradicting or approving* of actions or affections (as false or true), because they cannot be either. See Sayre-McCord (2008), Rachel Cohon (2008), and Elizabeth Radcliffe (2018) for discussion of these arguments.

which makes motives, volitions, and actions ineligible for contradiction and approval by reason but not by morality.[4]

This contrast between reason and morality reflects, Hume observes, a distinction within philosophy between "the speculative and the practical" and in life between what is inactive, reason, and what is active, "conscience, or a sense of morals" (*Treatise* 3.1.1.10, SBN 458). Morality, unlike reason, can and does cause or prevent actions by contradicting or approving of them – that is, by finding them vicious or virtuous, just or unjust, contrary to, or in accord with, duty, and obligation. Experience, Hume notes, teaches "that men are often govern'd by their duties, and are deter'd from some actions by the opinion of injustice, and impell'd to others by that of obligation" (*Treatise* 3.1.1.5, SBN 457).

Importantly, Hume does not think that morality always has this effect, just that it can have this effect. "'Tis one thing to know virtue," Hume points out, "and another to conform the will to it" (*Treatise* 3.1.1.22, SBN 465–6). What makes the difference to whether our opinions, moral or otherwise, influence our will and so shape our actions, turns on our sentiments and specifically on what we are interested in or concerned by.

To a large extent, Hume treats something as being of concern to us as it being the object of a desire, affection, or sentiment of ours. And he tends to treat such things as directly discoverable by us through introspection. But he explicitly recognizes that our desires, affections, and sentiments are sometimes so calm (to use his term) that they "are more known by their effects than by the immediate feeling or sensation" and are often just a matter of tendencies and dispositions (*Treatise* 2.3.3.8, SBN 417).[5] Hume goes on to warn that

> When any of these passions are calm, and cause no disorder in the soul, they are very readily taken for the determinations of reason, and are suppos'd to proceed from the same faculty, with that, which judges of truth and falshood.

[4] Missing that the argument depends crucially on Hume's account of what it takes for reason to contradict or approve of something, people have ended up finding in these arguments evidence that Hume holds that moral judgments necessarily motivate – a view often called internalism – while other judgments, specifically those that are a product of reason, do so only in conjunction with an appropriate desire, passion, or affection. But Hume is clear that moral judgments do not always motivate and that when they do, it is because of a concern for morality that someone might well fail to have. The problem with the sensible knave is not that he fails to recognize iniquity but that he doesn't care to avoid it except when personal advantage is in the offing (*Enquiry* 9.22, SBN 282–3).

[5] The switch to seeing the calm passions as tendencies and dispositions moves Hume toward a generally functionalist account of these passions and means he needs to offer some account of why we need to postulate such things, absent introspective evidence. The overarching grounds for doing so can be found in his view that "when in any instance we find our expectation to be disappointed, we must conclude that this irregularity proceeds from some difference in the causes" (*Treatise* 1.3.15.8, SBN 174), although a lot of work would need to be done either to defend the functionalist view or to identify the "difference in causes" that matters as passions.

(*Treatise* 2.3.3.8, SBN 417)[6]

Far from being determinations of reason, though, the presence or absence of these passions, as well as their objects and strengths, are what explain when and why the determinations of reason – concerning what is true or false – influence our wills.[7] For each and every such determination, people might be completely indifferent; whether they are is, Hume maintains, a matter not settled by the operations of reason but by the presence or absence of relevant passions or dispositions.

Some have thought that the very fact that Hume holds that reason needs to be supplemented by sentiment or passion in order to result in action precludes his believing in practical reason. After all, they point out, on Hume's view – and indeed, in his own words – "reason is perfectly inert" (*Treatise* 3.1.1.8, SBN 457–8). Isn't this a view according to which reason is simply not practical?[8]

Well, on Hume's view, neither reason nor sentiment *alone* results in action, even as either might happen to cause certain behaviors.[9] So neither, taken alone, in this way of thinking, is practical, even as together they are. Still, when reason is not alone, it can be practical and often is. What sentiment adds to reason is a concern for what reason discovers.

Strikingly, Hume's views concerning sentiment's role in explaining the practicality of reason have a remarkable echo in Kant's appeal to respect for the moral law.[10]

---

[6] Hume sees the calm passions as being of two kinds: "either certain instincts originally implanted in our natures, such as benevolence and resentment, the love of life, and kindness to children; or the general appetite to good, and aversion to evil, consider'd merely as such" (*Treatise* 2.3.3.8, SBN 417).

[7] Perhaps it is worth noting that admitting this role for the passions leaves completely open whether, when action occurs, the relevant passions need have been present before or independently of the products of reason with which they combine to produce action. Hume's position here is fully compatible with thinking that often reason itself causes the passions that, in combination with beliefs, gives rise to actions. What is ruled out is that reason causes these passions by discovering them to be conformable (or contrary) to reason. While such passions might be caused by reason, they are not themselves within the ambit of reason.

[8] See Jonathan Harrison (1976), Barry Stroud (1977), J. L. Mackie (1980).

[9] This is not to deny that sentiments cannot alone cause behavior, in contrast with causing an action. A sharp feeling of pain may well cause a grimace, for instance, but the grimacing in question, precisely because it does not depend on the person's understanding of her situation, does not count as an action she has performed. There is cause but not, to use Hume's term, a motive for the behavior.

[10] See "On the Incentives of Pure Practical Reason" in Kant (1788). In Kant's view, respect for the moral law works as an incentive, which is to say as "a subjective determining ground of a will whose reason does not by its nature necessarily conform with the objective law" ([72], p. 74). That is not to say, though, that the moral law depends for its effectiveness on any prior feeling. Rather, Kant maintains, the consciousness of the moral law itself gives rise the moral feeling of respect for the law and thus works not just as the "objective determining ground of the objects of action" but also as "a subjective ground of determination" [[75], p. 78). Aspects of Kant's view, as he emphasizes, turn on his rationalism concerning the moral law and on his conception of the free will, both of which Hume rejects. Yet his search for a feeling to let the law serve as the subjective ground of determination for the will reflects a sensitivity to the concerns that led Hume to hold that something (a feeling of some sort) must be added to recognition

Hume and Kant each recognize that people who hold the same beliefs, whether about morality or otherwise, might fail to will accordingly.[11] Each also appeals to something distinct from the moral belief or judgment to explain its impact (when it has an impact) on the will. For Hume, it is a concern for or interest in virtue; for Kant, it is a feeling of respect for the moral law.[12] Moreover, they agree that the relevant motivating feeling can be caused directly by recognition of what virtue or duty requires, unmediated by a separate feeling or desire. As Hume notes, "we have naturally no real or universal motive for observing the laws of equity, but the very equity and merit of that observance" (*Treatise* 3.2.1.17, SBN 483).[13]

At the same time, there are plenty of deep differences. Kant holds, for instance, that the beliefs in question are knowable *a priori*, and he holds that the moral feeling of respect has its source in our noumenal selves. Hume, in contrast, would reject both claims.[14] Moreover, of course, they hold dramatically different views of what sets the authoritative standard, with Kant defending the categorical imperative as the overarching standard of practical reason and Hume appealing instead to the informed and impartial reactions of those who take up the General Point of View.

Another difference, which is directly relevant to our concerns here, is that Kant counts respect for the moral law as a part of reason and its absence as a failure to be rational, whereas Hume works with a more restricted conception of reason and its requirements (as noted previously) – one that limits reason's constituent capacities to deductive and probabilistic reasoning and its scope of demands to what can be true or false. Yet this difference is merely one of taxonomy, neither deep nor principled. Kant does not hold that respect for the law is in accord with reason *because such respect is true*, nor that its absence is contrary to reason *because not feeling it is false*. Hume

---

of one's duty to explain when and why that recognition determines the will. See Guyer (2012).

[11] Aristotle, another famous defender of practical reason, similarly holds that the reason or intellect needs to be supplemented by something along the lines of what Hume regards as a passion, noting that "a faculty of practical thought is truth in agreement with the correct desire." He goes on to say that "the origin of action – in terms of the source of the movement, not its end – is decision, while that of decision is desire and rational reference to an end," "Thought by itself sets nothing in motion," and that decision is "either intelligence qualified by desire or desire qualified by thought" (Nicomachean Ethics, Book VI, 1139a 30–35, Rowe trans.). See Aristotle (2002).

[12] Exactly how Kant thinks of respect for the law, whether, for instance, it is bound up with recognizing the authority of the law, or is a separate feeling caused by that recognition, is a matter of dispute. What is clear (more or less) is that Kant holds that, absent such respect, an agent capable of acting from duty will fail to do so. See Reath (1989) and McCarty (1993).

[13] Hume makes this observation in the context of setting up a puzzle concerning how the laws of equity come to count as virtuous. The problem is that, in Hume's view (for reasons that needn't concern us), each virtue must be such that there is some original motive to act as the virtue requires that is independent of a recognition of it being a virtue, yet the laws of equity *seem* not to meet this requirement. Hume's solution is to distinguish natural from artificial motives and argue that the original motive to equity, which is independent of a recognition of equity as a virtue, is a motive that depends on artifice, not nature.

[14] More generally, Kant allows the possibility of *a priori* synthetic truths and transcendental proofs, neither of which are in Hume's repertoire.

meanwhile does not rule out that the authoritative standards might require, among other things, a concern with virtue (which might lead a practically virtuous person who sees something as virtuous to act accordingly). In fact, for all that Hume argues, he could consistently embrace a broader conception of reason according to which a motive (to act as virtue or duty requires) is part and parcel of having reason.[15] What is important for our purposes is that such an expanded account of reason – call it "practical reason" – leaves all the key elements and arguments of Hume's view (which appeal to the "strict and philosophical" sense) in place, even as it adds in *a concern to act as virtue requires.*[16]

In whichever way the capacities crucial to the practicality of practical deliberation are labeled, Hume and Kant (and many others who hold theories of practical reason) share the view that success or failure to act as the standards of practical reason require involves not just the capacity to make judgments (concerning virtue or duty) but the presence of sentiments, feelings, or motives that are not themselves matters of judgment (even if they are required by the relevant standards). Some such extra element is needed so long as it is possible, as Hume and Kant both acknowledge it is, for someone to make the relevant judgment and yet fail to act accordingly. As a result, it is a mistake to think that just because reason, as Hume conceives it, is not sufficient alone for action, the view he develops should not be counted as a view of practical reason.

In this section, I have concentrated on Hume's insistence that sentiment, no less than reason, is central to practical deliberation being practical. Absent the relevant sentiments, Hume insists, whatever judgments reason might lead us to make, they will remain inert.[17] The concern has been, in effect, with what Hume calls the influencing motives of the will, of which he considers reason and sentiment equally required.

Hume is similarly convinced that reason and sentiment are equally required in order to understand *practical deliberation.* At bottom, he thinks that in order to understand our taking considerations as counting in favor of (or against) things, we need to pay attention to the nature of approbation and disapprobation.[18] I turn to this aspect of Hume's account in the next section.

---

[15] Hume doesn't actually hold such a view, but for the substantive reason that if one's moral views are wrong, it might not be virtuous to act as they require. Kant seems to have more confidence that people can and do successfully recognize their duty and so sees acting accordingly as unproblematic.

[16] In Hume's "strict and philosophical" sense of reason, its exercise already supposes the dispositions or concerns that lead one to make deductive and inductive inferences and form beliefs accordingly. He could, without any serious change, simply add in to his account an understanding of the exercise of "practical reason" as involving the dispositions or concerns that lead one to act as practical reason's standards (as set by the General Point of View) require.

[17] In insisting on this, I have just argued, Hume is advancing a view that Kant in effect acknowledges when he introduces "respect for the law" as a necessary part of the story of when and why human beings who are rational beings who can recognize their duty can succeed not merely in acting in accord with duty but in acting from duty.

[18] While I will not go into the point here, I think Hume sees approbation as playing the very same

role in explaining when and why we have reasons to believe as in explaining when and why we have reasons to do or choose various things.

Hume on the activity of practical deliberation

According to a familiar instrumentalist picture, practical deliberation is a matter of figuring out the effective means to satisfying our desires and then acting accordingly. Our desires set our goals and constrain what we are willing to do to achieve them; our reason then canvasses the options available within those constraints, working to identify what will maximize their satisfaction. Hume clearly does think we engage in this sort of deliberation, and he sees the results as having a seriously practical impact on how we act. Reason might here be "the slave of the passions" (*Treatise* 2.3.3.4, SBN 414–5), but like so many slaves, it does a tremendous amount of important work.

Yet to think of Hume's account of the activity of practical deliberation entirely in terms of means-ends reasoning flattens the terrain dramatically and misses the ways in which Hume recognizes that, and offers the resources to explain how, we can deliberate about ends. It also obscures completely his account of the extent to which, even in deliberating about means, we can be, and often are, concerned with identifying and acting in light of what we take to be reasons – that is, considerations that count in favor of or against certain courses of action – regardless of our desires.

Appreciating Hume's richer story requires paying attention to the distinction he draws between direct and indirect passions, which then reveals the extent to which he was sensitive both to the reactive attitudes and to attitudes that are "reason-responsive." With those resources on board, I will argue, Hume is able to offer an account of the activity of practical deliberation that distinguishes

> (i) considerations *influencing* our attitudes, from
>
> (ii) *our thinking* that those considerations count in favor (or against) those attitudes, from
>
> (iii) those considerations *actually counting in favor (or against)* those attitudes (whether or not we recognize that they do and whether or not they have an influence).

Early on in Book II of the *Treatise*, Hume distinguishes ideas from impressions, then distinguishes original from secondary impressions, then distinguishes, among secondary impressions, those that are direct from those that are indirect (*Treatise* 2.1.1.4, SBN 276–7). In discussing the indirect passions, he focuses on love, hate, pride, and humility,

developing an intriguing and complex account of when and why we feel these passions and the extent to which they are felt in light of, and on the basis of, considerations that weigh in favor of approving or disapproving of those who are the objects of these passions (ourselves, in the case of pride and humility; others, in the case of love and hate).[1]

As Hume makes clear, the story of these four passions extends to the more general sentiments of approbation and disapprobation, which are at the heart of how Hume makes sense of practical deliberation. Approbation and disapprobation, he maintains, are "nothing but a fainter and more imperceptible love or hatred" (*Treatise* 3.3.5.1, SBN 614).[2] What they have in common is that they are felt towards their object only when, and because, one believes their object is related to something that has features consideration of which is appealing or off-putting.[3] If either the thing were discovered not to have the relevant feature, or, although it had the feature, that feature was neither appealing nor off-putting but instead a matter of indifference, in those cases, the approbation or disapprobation would, Hume maintains, disappear. That the object has the qualities in question serves as a consideration that, in effect, *weighs* in favor of approving it but only so long as those qualities are themselves appealing. The same story, in mirror image, goes for disapproval and the unappealing qualities that objects may have.[4]

The guiding idea is that the features in question, when considered by the person whose reactions we are thinking about, must be such that *that* person finds them appealing. That she does find them appealing, though, is often *not* among the considerations that end up weighing positively with her. Still, the fact that some consideration is appealing to her explains why that consideration weighs positively. Of course, the very

[1] (*Treatise* 2.1.2.1, SBN 277- T 2.1.5.11, SBN 289–90). The weighing in question is a matter of in fact inclining us one way or another when it comes to feeling approbation or disapprobation; it is not, of itself, a matter of the considerations counting in favor of those feelings nor a matter of the person who is doing the considering thinking that the count in favor.

[2] Whether they are themselves passions or not is a bit unclear. What is clear is that approbation and disapprobation are, like love and hate, indirect sentiments. One important difference is that love, hatred, pride, and humility all, in Hume's telling, have people as their objects, whereas approbation and disapprobation evidently take within their scope all manner of possible objects. That difference notwithstanding, they too are felt, when they are, only in light of their object's perceived relation to some feature that has appealing or off-putting features.

[3] Feeling positively about some thing, in light and because of some considerations, is, in Hume's view, just to approve of it.

[4] Hume's explanation of the indirect passions is so complex that people commonly regard it as too baroque to be worth taking seriously. As a result, it has received relatively little attention. My own view is that the complexities in play are no more than the phenomena to be explained require. Fortunately, though, for our purposes here, we do not need to go into the details. For a discussion that does take Hume's explanation seriously, see Pall S. Ardal (1966). See my (2013) for a description of the explanation Hume offers.

idea that one finds something appealing might, itself, be unappealing in a way that leads to disapproving of oneself for finding such things appealing.[5]

Significantly, although approval and disapproval are felt only because we find some features appealing, they are not felt as a means to, nor for the sake of, our desires or ends, nor even as a means to, nor for the sake of, the pleasing feeling of approval. The approval, when felt, is independent of our desires or ends or interest in feeling pleasures, even as it is itself a pleasant response to what we are considering. This is true even though the favorable light in which we see the features of an object often leads us to approve of what has, or might acquire, those features.[6] In such cases, the approval to which the considerations give rise might lead us to promote the qualities in question. Yet we do not feel the approval because the approval has that effect.

When the features in question are seen in a favorable or unfavorable light, they weigh with us in determining what we approve or disapprove of – thus they serve as considerations that *weigh* in favor of or against approving whatever we might be considering, including courses of actions and ways of responding to the actions of others. This is, importantly, not the same as our seeing those considerations as *counting in favor* of approving of them.

So far, we are simply talking about people feeling approval (or not) in light of various considerations. We get to potentially practical *deliberation*, though, once we move to cases in which people see themselves as having various options and are thinking about which to take, letting what they in fact approve of settle what they will do. In these cases, a person will be taking considerations into account that are, in fact, weighing with her in favor of or against the various options, often in ways that mean there are a lot of considerations in play, weighing more or less strongly for or against different options. To come to a conclusion, in this context, is to end up, all told, approving of one option over the others. Such deliberation will be practically effective when the process of weighing the considerations leads one to take the option one (most) approves of, because one approves of it.[7]

This is all, as Hume sees things, something that can happen without anyone *thinking* that the considerations *count in favor of* (or *count against*) the objects of our approval or disapproval. The considerations might weigh with us without being seen by us as providing reasons for our attitudes. All the considerations are doing is weighing, one way or another, thanks to their role in causing approbation or disapprobation. They

[5] Plato makes a lot of this with the story of Leontius being disgusted by his desire to gaze upon corpses. See *The Republic* (ca 380 bc), Book IV, 439a.

[6] The 'often' here is important. It is no part of Hume's theory that we always approve of things that we see as having features that we find appealing. A lot can interrupt the movement from noticing that something has a certain feature that is appealing and approving of it (because of that feature).

[7] It is worth noting that the process of weighing need not be linear in its effects. While Hume does not go into this, his view is compatible with thinking, say, that considerations might interact in ways that mean one consideration that would normally have weighed with one might, in light of other considerations, not weigh at all or might interact in ways that mean two considerations weigh together more than the sum of each taken individually.

thus so far figure in the explanation, but not the justification, of our approving or disapproving as we do and of our acting as they would have us act.

Still, they put Hume in a position to contrast, as Aristotle does, merely voluntary action from voluntary actions performed in light of and because of practical deliberation. Both require that one be responding to one's beliefs concerning one's circumstances, but the latter involves canvassing (what one takes to be) one's options and choosing among them in light of considerations that on balance weigh in favor of one or another.[8]

Yet Hume recognizes that we can and do distinguish, in effect, between (i) approving or disapproving of something and (ii) that thing being approvable (that is, as meriting the approval or disapproval it might receive).[9] Something – a character trait, an action, an institution, and so on – is approvable, he argues, thanks to it being such that it would be approved of by those who take up the General Point of View (and so are appropriately informed, impartial, and sympathetic to all who are relevant). Once that standard is available, people are able to think of the considerations that might be inclining them to approve (or disapprove) of various options *as reasons*, that is, as considerations in light of which things might merit approval (or not).

Importantly, people might find themselves approving or disapproving of people, or courses of action, or the reactions of others, in light of various considerations, even as they themselves recognize that those considerations are not reasons for the approval or disapproval they feel. They might for instance recognize that they are racist, or sexist, or classist in ways that have an impact on their attitudes and actions while thinking the considerations that in fact weigh with them should not have that impact.

With the distinction between, on the one hand, being moved by a consideration and, on the other hand, thinking of it as counting in favor of something – that is, thinking of it as a reason – Hume is in a position to explain how, in engaging in practical deliberation, we are able not only to think about means to our ends but also about whether the ends we find ourselves with are worth pursing (i.e., approvable) or not. Considering which ends are worth pursuing, on Hume's account, involves reflecting not (solely) on how we actually feel or what we currently desire but on what we would approve of from the General Point of View.[10] Even when we cannot actually take up

[8] Aristotle's account of the distinction between merely voluntary actions and voluntary actions performed in light and because of deliberation is detailed in Book III, Chapters 1–4, of the *Nicomachean Ethics*.

[9] Hume's discussion is usually in terms of the virtue or merit of something rather than it being approvable, but the contrast in play is the same. Smith takes up the very same contrast, framing it specifically in terms of the difference between securing praise and being praiseworthy, noting that we desire "not only praise, but praise-worthiness; or to be that thing which, though it should be praised by nobody, is, however, the natural and proper object of praise" (Smith (1790), 114). In Smith's account, whether something is praiseworthy turns on whether it would be approved of by an Ideal Spectator who is an "impartial and well-informed spectator" (Smith (1790), 130).

[10] That one has certain desires or ends may of course be among the considerations that count in favor of the value of possible ends; the point here is that they do not, in Hume's account, at all settle whether an end is worth pursuing.

that point of view, or do not in fact feel as we would if we were to take it up, we can nonetheless *think* in terms of right and wrong, good and bad, and virtue and vice, and sometimes act accordingly. In acquiring these reflective and cognitive resources, people become what Kant characterizes as "rational agents" – agents able not merely to conform to laws but to act according to their "conception of the law".[11]

Thus, on Hume's account, in deliberating about our options, we can and do take into account not merely considerations concerning how we might satisfy our desires or achieve our ends but also considerations concerning the value, permissibility, and virtue (or otherwise) of the courses of action we might take and the ends we might adopt. And we can and often do reach conclusions that, by approving or contradicting the options, lead us to act or refrain from acting. Whether our conclusions have this effect depends, of course, on our being concerned with value, permissibility, and virtue (which virtue itself would normally require).

Thus, Hume makes important room for our deliberating about what to do specifically in terms of what is valuable, permissible, or virtuous and then acting accordingly as a result. Yet Hume's account allows that other considerations that might weigh with a person can count *as reasons* for or against the actions she is considering, even if she is not thinking of them as reasons (either because she lacks the relevant concepts or because her attention is on something else). These will be considerations the weighing of which, with the person, would secure approval from those taking up the General Point of View. Success in having one's practical deliberation appropriately sensitive to the reasons one has does not, on Hume's account, require that one be conceiving of those considerations as reasons. What matters is that the right considerations weigh in the right way.[12]

The end result is a rich account of what we are doing in engaging in practical deliberation. In the first instance, we are thinking about (what we take to be) our options, in light of what we believe about them, with these considerations weighing with us, positively, negatively, or not at all, in determining what we approve of doing in a way that leads to action.

Among the considerations available are whether the options are, in fact, approvable: whether they are right or wrong, good or bad, virtuous or vicious. Some people, needless to say, think about what to approve of without ever really thinking about what is approvable. Still, in weighing various considerations they are engaging in practical

[11] "Everything in nature acts according to laws. Only a rational being," Kant observes, "has the power to act according to his conception of laws, i.e., according to principles, and thereby has he a will" (Kant (1785), 23, AK 4:412).

[12] In Sayre-McCord (1994), I argue that, according to Hume, it is vitally important that we be able to access the General Point of View; otherwise, having the standard would not work to solve the problem the solving of which is the *raison d'être* for having a standard of approvability. Other people, of course, have other views of Hume. And, independently of Hume, people (like Firth and Smith) defend views that are, in the ways that matter to this chapter, like Hume's while holding that their equivalent of the General Point of View may be wholly inaccessible.

deliberation (though they are considering what to do without being concerned with whether doing it is good or right or justified). And, depending on the considerations they take into account and how those considerations weigh with them, such people may well succeed in approving what is approvable and do so in light of the relevant reasons (even though, by hypothesis, they are not thinking of them as reasons). Also, though, other people might concentrate in their deliberations on what is approvable, yet be influenced by considerations that do not in fact count in favor or against the options they consider in ways that lead them to the wrong conclusions or to the right conclusions but not for the right reasons.

Which ways of deliberating, with which aims in mind, count as the right way of engaging in practical deliberation turns, in Hume's view, on what would be approved of from the General Point of View. He takes more or less clear stands when it comes to certain contexts, regarding, for instance, the various virtues he discusses – as long as their respective demands are not in tension. But, when it comes to practical deliberation, he develops nothing like his "Rules by which to judge of cause and effects,"[13] although he does in several places refer to 'moral precepts' and recommends relying on them to "fortify the mind against the illusions of passion."[14] It is unclear how exactly Hume would think of developing rules for practical deliberation. Presumably, though, he would endorse whatever rules might be developed to the extent relying on them, or conforming to them, would secure approval from the General Point of View. These rules might well include the standard principles of decision and game theory, on the grounds that reasoning in accord with them worked to promote optimal outcomes, along with other principles that have been advanced as "rational principles" that require not just that in willing an end we will the means (or abandon the end) but also, say, that we avoid weakness of will, either by acting (or intending to act) as we judge we should, or by abandoning the view that we should so act, and others as well.[15] However such rules might be developed, though, Hume is committed to holding that such principles cannot be defended as standards for us simply by pointing to analytic truths in which they might figure nor by arguing that flaunting them involves believing or doing anything that is itself false.

[13] *Treatise* (1739–1740), BK I, Pt. III, Sect. XV.

[14] "The Sceptic" (1742), p. 179.

[15] See Broome (1999), Broome (2001), and Kolodny (2005) for divergent views on how to understand such rules.

In sum

I have not here tried to defend what I take to be Hume's theory of practical reason. There are plenty of worries and objections that might be raised and that, as I see it, constitute serious challenges. I have, though, done what I can to make the case for thinking there is such a theory and that it offers plausible accounts of what we are doing when we engage in practical deliberation, of what capacities are required for that activity to be effective, and of the standards that matter when engaging in it.

Admittedly, in doing this, I have pressed against the many standard and influential interpretations that see – and sometimes celebrate – Hume as either a skeptic or an instrumentalist. While I do this without apology, I am mindful that the need to do this itself constitutes reason to be suspicious of the interpretation I offer. Right now, though, I see the standard interpretations, whatever the advantages of the positions they identify, as missing deep and interesting aspects of Hume's actual view, aspects that suggest a richer, and more robust, theory of practical reason than is usually appreciated.

References

Ardal, Pall S. (1966). *Passion and Value in Hume's Treatise* (Edinburgh: University of Edinburgh Press).

Aristotle (2002) *The Nicomachean Ethics*, trans. Christopher Rowe, with introduction and commentary by Sarah Broadie (Oxford: Oxford University Press).

Boyle, Deborah (2012) "The Ways of the Wise: Hume's Rules of Causal Reasoning," *Hume Studies*, Vol. 38, No. 2, November, pp. 157–182.

Broome, John (1999) "Normative Requirements," *Ratio*, Vol. 12, pp. 398–419.

Broome, John (2001) "Normative Practical Reasoning," *Proceedings of the Aristotelian Society, Supplementary*, Vol. 75, pp. 175–193.

Cohon, Rachel (1997) "Is Hume a Noncognitivist in the Motivation Argument?" *Philosophical Studies*, Vol. 85, pp. 251–266.

Cohon, Rachel (2008) *Hume's Morality* (Oxford: Oxford University Press).

Firth, Roderick (1951). "Ethical Absolutism and the Ideal Observer," *Philosophy and Phenomenological Research*, Vol. 12, No. 3, pp. 317–345.

Guyer, Paul (2012) "Passion for Reason: Hume, Kant, and the Motivation for Morality," *Presidential Address, Proceedings and Addresses of the American Philosophical Association*, Vol. 86, No. 2, pp. 4–21.

Hampton, Jean. (1995) "Does Hume Have an Instrumental Conception of Practical Reason?" *Hume Studies*, Vol. 21, No. 1, pp. 57–74.

Harrison, Jonathan (1976) *Hume's Moral Epistemology* (Oxford: Oxford University Press).

Hubin, Donald C. (1999) "What's Special About Humeanism?" *Nous*, Vol. 33, No. 1, pp. 30–45.

Hume, David (1739–1740) *A Treatise of Human Nature*, ed. L. A. Selby-Bigge, revised by P. H. Nidditch (Oxford: Oxford University Press, 1978).

Hume, David (1742) "The Sceptic" originally published in Essays, Moral and Political, vol. II, reprinted in Essays: Moral, Political and Literary, Eugene F. Miller (ed.), (Indianapolis: LibertyClassics).

Hume, David (1751) *An Enquiry Concerning the Principles of Morals*, in Enquiries, L. A. Selby-Bigge (ed.) revised by P. H. Nidditch (Oxford: Clarendon, 1975).

Hutcheson, Francis (1742) *Illustrations Upon the Moral Sense*, ed. Aaron Garrett (Indianapolis: Liberty Fund, 2002).

Kant, Immanuel (1785) *Grounding for the Metaphysics of Morals*, trans. James W. Ellington (Indianapolis: Hackett Publishing Co., 1993).
Kant, Immanuel (1788) *Critique of Practical Reason*, trans. L. W. Beck (New York: Macmillan, 1985).
Kolodny, Niko (2005) "Why Be Rational?" *Mind*, Vol. 114, No. 455, pp. 509–563.
Korsgaard, Christine (1986) "Skepticism About Practical Reason," *Journal of Philosophy*, Vol. 83, No. 1, pp. 5–25.
Lewis, David (1988) "Desire as Belief," *Mind*, Vol. 97, pp. 323–332.
Locke, John (1689) An Essay Concerning Human Understanding. An Essay Concerning Human Understanding*, ed. Peter H. Nidditch (Oxford: Clarendon Press, 1975).*
Mackie, J. L. (1980) *Hume's Moral Theory* (London: Routledge).
McCarty, Richard (1993) "Kantian Moral Motivation and the Feeling of Respect," *Journal of the History of Philosophy*, Vol. 31, No. 3, pp. 421–435.
Millgram, Elijah (1995) "Was Hume a Humean?" *Hume Studies*, Vol. 21, No. 1, pp. 75–93.
Plato. *The Republic* (ca 380 bc), trans. G. M. A. Grube, revised by C. D. C. Reeve (Indianapolis: Hackett, 1992).
Radcliffe, Elizabeth (1997) "Kantian Tunes on a Humean Instrument: Why Hume Is Not Really a Skeptic About Practical Reason," *Canadian Journal of Philosophy*, Vol. 27, No. 2, pp. 247–269.
Radcliffe, Elizabeth (2018) *Hume, Passion, and Action* (Oxford: Oxford University Press).
Reath, Andrews (1989) "Kant's Theory of Moral Sensibility: Respect for the Moral Law and the Influence of Inclination," *Kant-Studien*, Vol. 80, pp. 284–302.
Sayre-McCord, Geoffrey (1994) "On Why Hume's 'General Point of View' Isn't Ideal – and Shouldn't Be," *Social Philosophy and Policy*, Vol. 11, No. 1, pp. 202–228.
Sayre-McCord, Geoffrey (1996) "Hume and the Bauhaus Theory of Ethics," *Midwest Studies in Philosophy*, Vol. XX (Notre Dame: University of Notre Dame Press), pp. 280–298.
Sayre-McCord, Geoffrey (2008) "Hume on Practical Morality and Inert Reason," in *Oxford Studies in Metaethics*, edited by Russ Shafer-Landau (Oxford: Oxford University Press), pp. 299–320.
Sayre-McCord, Geoffrey (2013) "Hume and Smith on Sympathy, Approbation, and Moral Judgment," *Social Philosophy and Policy*, Vol. 30, No. #1–2, pp. 208–236.
Schafer, Karl (2015) "Hume and Practical Reason: Against the Normative Authority of Reason," in *The Oxford Handbook of David Hume*, edited by P. Russell (Oxford: Oxford University Press).
Setiya, Kieren (2004) "Hume on Practical Reason, Philosophical Perspectives," *Ethics*, Vol. 18, pp. 365–389.
Smith, Adam (1790) *A Theory of Moral Sentiments*. Sixth Edition. (London: A. Millar).
Smith, Michael (1987) "The Humean Theory of Motivation," *Mind*, Vol. 97, pp. 36–61.
Smith, Michael (1994) *The Moral Problem* (Oxford: Blackwell).

Stroud, Barry (1977) *Hume* (London: Routledge & Kegan Paul).
Wollaston, William (1722) *The Religion of Nature Delineated*. Eighth Edition. (London: Printed for J. Beecroft, J. Rivington, J. Ward, R. Baldwin, W. Johnston, S. Crowder, P. Davey and B. Law, and G. Keith, 1759).

10. Kant's approach to the theory of human agency

I Introduction

Any theory of practical reasoning implicitly or explicitly relies on some characterization of the practical reasoner. Who are we, as creatures who employ practical reasoning? What motivational capacities do we have? What way of exercising these capacities counts as practical reasoning? Let's call these questions about the nature of human agency. My concern here is about how philosophers go about discussing this topic.

In particular, I am interested in the difference between a broadly mechanistic tradition, exemplified by belief-desire psychology and its variants, and a Kantian tradition. The mechanistic tradition describes human agency as a complex of mental states and mental events. It identifies agency with a psychological process in which, for example, beliefs and desires combine "in the right way" to produce behavior (Davidson 1963). You have a desire for a cold drink, and a belief that by opening the refrigerator, you will get a cold drink. When these mental states interact in the right way, the resulting behavior amounts to an exercise of your agency.

The Kantian tradition characterizes human agency in very different terms. Kant writes, "Everything in nature works according to laws. Only a rational being has the capacity to act *in accordance with the representation* of laws, that is, in accordance with principles, or has a *will*" (Kant 1996a: 66/4:412). Kant does not try to analyze this capacity in terms of an underlying mechanistic process. Rather, he seems to conceive of the will as something that stands over and above all of the incentives that influence it (Kant 1996b: 49/6:23–24; Allison 1990: 189). When you have a desire to drink a cold drink, Kant would say, you face a choice about whether to act on that desire. You have to exercise your will in relation to your desire. To do that, Kant maintains, is to make a "principle" the "ground" of the "determination" of your will (Kant 1996a: 66/4:413ff.; 153/5:19ff.).

Most contemporary action theorists tend to be puzzled by Kant's way of characterizing motivation. By and large, they work within the mechanistic tradition. Granted, that tradition is by no means monolithic. There are lively internal debates, for example, about which conceptual vocabulary to use. Are beliefs and desires the basic, irreducible mental states, or are there more complex states not reducible to these, such as intentions or volitions? Should mental states be described in folk psychological or more scientifically precise terms? Should they be individuated in functional or causal terms? These debates notwithstanding, philosophers in this tradition find it difficult to understand and take seriously Kant's conceptual vocabulary. This in turn makes it hard for them to see how Kant is offering a theory of human agency at all. They

have no trouble seeing how Kant puts forth a normative theory, a theory of rational or morally worthy action. But they find it harder to see how he is putting forward a descriptive account of what human agency *is*.

I want to promote a more fruitful dialogue on this front. I will argue that in order for the mechanist to understand what the Kantian is up to, it is essential to articulate more clearly Kant's philosophical method. Doing this reveals that the two traditions frame the philosophical question about agency differently. The mechanist's theory is shaped in large part by the method of natural science. It addresses the question, "What happens when someone acts?" (Velleman 1992).[1] Kant's theory, by contrast, is shaped by his distinctive method of critique. It addresses the question, "What am I doing insofar as I am acting?" As long as this difference remains unarticulated and unexplained, genuine dialogue between the two traditions will tend to stall.

I will proceed as follows. First, I will make the dialectic I am interested in more vivid by citing two illustrative stretches of text. Next, I will identify exactly why the mechanist finds the Kantian position puzzling. Finally, I will explain Kant's method of critique and show how it can provide the basis for a response to the mechanist's worries. My aim here is not to argue that Kant's approach is better. I merely want to show that it constitutes a substantive and philosophically well motivated alternative.

[1] Velleman rejects a standard way of answering this question, but he does not reject the question.

II Two dialogues

Since my topic is the way one tradition finds another tradition puzzling, it would be helpful to start with a passage expressing that puzzlement. Such passages are surprisingly hard to find. Although mechanists and Kantians characterize motivation in very different terms, they rarely talk to each other about that fact. Kant's own concerns about mechanistic accounts of agency appear in the context of his discussion of transcendental freedom (Kant 1996a: 216ff./5:96ff). That will become relevant later, but it does not serve as an accessible starting point. Instead I will use two stretches of debate to which Kant himself is not a party, in order to illustrate the dialectic I am interested in. One is an eighteenth-century exchange between Leibniz and Clarke, and the other is a twenty-first-century dialogue between Michael Bratman and Christine Korsgaard. Clarke and Korsgaard will serve as Kant's surrogates in these passages. Korsgaard works self-consciously within the Kantian tradition. Clarke predated Kant, and Kant would have criticized his overall view as a form of dogmatic rationalism. But in this passage, he sounds similar enough to Kant for that difference to be irrelevant. My aim in looking at these texts is simply to focus the reader on how each exchange leaves the mechanist puzzled.

Leibniz and Clarke

Although Leibniz and Clarke were both in some sense rationalists, they disagreed about how to characterize agency. Leibniz compares an agent to a balance, suggesting that an agent's motives determine his actions in the way that weights on the trays determine the movement of the scale (Leibniz and Clarke 2000: 7/L II 1). Clarke objects to the analogy. In his words,

> A balance is not an agent, i.e. doesn't act, but is merely passive and acted on by the weights; so that when the weights are equal, nothing moves it. But thinking beings are agents; they aren't passive things that are moved by their motives as a balance is moved by weights; rather, they have active powers through which they move themselves, sometimes upon the view of strong motives, sometimes upon weak ones, and sometimes where things are absolutely indifferent.
>
> (Leibniz and Clarke 2000: 29/C IV 1–2)

An agent, Clarke maintains, is a creature who has the "active power" of self-movement. Self-movement in this agential sense, he argues, cannot in principle be identified with movement caused by an imbalance of forces. Now it might seem that Leibniz embraces something closer to Clarke's picture when he writes that motives "incline without necessitating" (Leibniz and Clarke 2000: 37/L V 9). But Leibniz also holds that we cannot but affirm our strongest motives. Clarke reads this as implying that the will is nothing more than a further cog in the motivational machine. As Kant himself will later write, this Leibnizian freedom is "nothing better than the freedom of a turnspit, which, when once it is wound up, also accomplishes its movements of itself" (Kant 1996a: 218/5:97). True agency, Clarke argues, is a power to act on the stronger or the weaker motive. It is thus a power we exercise independent of the pressure exerted by the motivational machinery, albeit with that machinery in view.

Leibniz is unpersuaded. He replies:

> strictly speaking motives don't act on the mind in the way weights act on a balance. What really happens is that the mind acts by virtue of its motives, which are its dispositions to act. And so to claim as Clarke does here that the mind sometimes prefers weak motives to strong ones, and even that it sometimes gives its preference to something that is indifferent, putting that ahead of any motives – this is to divide the mind from the motives, as though they were outside the mind and distinct from it as the weights are distinct from the balance, and as though the mind had, as well as motives, other dispositions to act, by virtue of which it could accept or reject the motives. Whereas the motives include all the dispositions that the mind can have to act voluntarily – not only its reasons but also any inclinations it has because of passions or other preceding impressions.
>
> (Leibniz and Clarke 2000: 38–9/L V 15)

According to Leibniz, Clarke fails to understand that, in the mechanistic account, our motives do not push us around. Rather, they *are* us. When our dispositions – intellectual and sensible – move us, we count as moving ourselves. There is no reason to "divide the mind from the motives." Doing so only begs the further question of what the mind-behind-the-mind is, and what motivates its activity.

Bratman and Korsgaard

Now fast forward to the twenty-first century. Here again, two prominent philosophers of agency find themselves in a version of the same debate. Michael Bratman, working within an empiricist tradition, builds upon the neo-Humean, belief-desire model of agency. He argues that what makes an event an action is that it is caused in the right way by a complex structure of attitudes that includes not only beliefs and desires

but also intentions (Bratman 1987, 2007). Korsgaard, in a commentary on Bratman's theory, asks whether this account "makes it possible for us to lay claim to our actions and attribute them to our active selves." As she sees it, Bratman merely stipulates that a certain mechanistic process counts as agency. She finds something arbitrary or unsatisfying in that move:

> We cannot just pick out some of those states and say "we are active when those are operative": we have to explain why that should be so.
>
> (Korsgaard 2014: 203)

Bratman would, of course, agree. He takes pains to offer a story about why the process he identifies should count as our activity. That story, as Korsgaard is aware, appeals to the idea that when intentions operate in the right way, they constitute the relations that make up the agent's identity, in a descriptive, Lockean sense. (Bratman 2007) According to Bratman, it is this connection to the agent's identity that makes certain processes count as the agent's activity.

But Korsgaard is not convinced. She denies that this account "adequately captures the element of self-determination ... because the Lockean notion of self-constitution requires only that the agent conform to his principles, not that he himself chooses them or stands in an active relation to them" (Korsgaard 2014: 201). The Lockean self, as she sees it, is just a further state of the mechanism, and Bratman's view is that a certain process counts as our activity because it results in, or realizes, a state of the mechanism that counts as our identity. But as Korsgaard sees it, this still amounts to nothing better than the freedom of a turnspit. The self-movement that is our agency cannot, in principle, be identifiable with the operation of mental states in us. Rather, our self-movement has to somehow be our doing, a way of operating upon our motives from a position outside them:

> For our mental states and attitudes to be expressions of our own activity, we must impose the forms of mental self-determination upon that activity by following the norms of mental activity.
>
> (Korsgaard 2014: 203)
>
> It is only insofar as we exercise this power – the power to "impose form" on given psychological material – that we "render ourselves the kind of active beings whose movements can be said to have their sources in the self"
>
> (Korsgaard 2014: 199).

Bratman is not convinced. He replies:

> On the planning theory, when there is relevant norm-guided activity grounded in appropriate plan structures, there is self-governed activity...

> . But there need be no further activity of the agent that makes the agent into a self-governed agent. In contrast, Korsgaard seems to interpret "render ourselves"/"make ourselves" as involving some further activity of the agent herself. And it is here that we get something that sounds like a homuncular agent. But why think self-governance requires this further form of making or rendering by you? This seems, as it were, one activity too many.
>
> (Bratman 2014: 326)

He continues,

> Korsgaard's theory interprets [self-determination] in a strong sense that goes beyond the modest metaphysics of the planning theory... . What we need is a model of self-governance in which there is not, behind it all, an agent inside pulling the strings.
>
> (Bratman 2014: 327)

What Bratman objects to in Korsgaard's theory is essentially what Leibniz objects to in Clarke's: namely that it "divides the mind from its motives." Either Korsgaard is simply begging the question by positing an agent behind the agent, or she is positing an obscure power of "agent-causation" – construed in Kant's terms as "form-imposition" – as a genuinely explanatory notion. Either way, Bratman does not feel the pull of a genuine challenge to his position.

III A question of method

To be clear, neither side is moved by the other at this point. And yet, I believe most contemporary readers would say that Kant's surrogates are in the weaker position. Why? I think there may be a shared perception that the only alternative to a mechanistic account of agency is an account that tacitly relies on an unacceptable conception of free will. I'll call this position "spooky libertarianism." Spooky libertarianism is a version of incompatibilism according to which we are indeed free, and our freedom consists in a distinctive capacity to intervene in the mechanism of nature from a standpoint outside that mechanism. This capacity, sometimes called "contra-causal freedom" or "agent-causation," is undetermined by prior events, but it can initiate new chains of events (Campbell 1951; Chisholm 1964). The spooky libertarian, as I will imagine her here, maintains that a full explanation of agential (and moral) attributability has to make reference not just to the mechanism of nature but also to this special, contra-causal power. Otherwise it cannot explain "how the agent gets into the act."[1]

I take it that when Bratman worries that Korsgaard's theory "goes beyond the modest metaphysics of the planning theory," he is concerned about the spookiness of spooky libertarianism. He worries that Korsgaard is at least tacitly committed to the idea that we produce our actions by exercising a real but non-empirical, and essentially mysterious, power of freedom. Leibniz, on the other hand, is certainly not worried about keeping his metaphysics modest. Unlike Bratman, he openly allows for "immaterial substances" and their agential powers to play a role in his mechanistic picture (Leibniz and Clarke 2000: 7/L II 1). As such, Leibniz's objection to Clarke's account is not that it is spooky but that it is circular. In "divid[ing] the mind from its motives," it purports to explain the agency of the individual human being by simply positing an individual human agent who stands behind the motives he has. Leibniz takes this to be question-begging. Bratman is also concerned about circularity in Korsgaard's account. But it is important to notice that the two worries are distinct. One is about metaphysical excess, and the other is about explanatory circularity.

The best way to defend Kant is to defend his picture of motivation against both worries. Now it might seem that the most direct way to do this is to focus on Kant's theory of freedom, with the aim of showing that transcendental freedom is in fact a metaphysically modest notion, unlike the spooky libertarian's contra-causal freedom. I believe there is at least some support for such an interpretation, and I will say a bit more about it later in the chapter. But that is not where I want to start. First

[1] Velleman (1992) uses this helpful phrase, but he does not endorse libertarianism.

I want to focus on the circularity objection, which I take to be equally important. Doing so will allow me to articulate the salient methodological differences, which often go unacknowledged, between the mechanistic approach to the theory of agency and Kant's own approach.

The mechanist sees his approach as avoiding circularity because it explains "what happens when someone acts," without appealing to the notion of action. As such, he tacitly assumes that the job of the theory of agency is to explain what happens when someone acts. Insofar as he charges Kant and his surrogates with explanatory circularity, he assumes they share this conception of what the job of the theory is. Presumably, they are trying to explain what happens when someone acts, but they do so only by making reference to someone acting.

Let me spell out the mechanist's approach in a bit more detail. Its main features are as follows:

1. The primary object of inquiry, the action of the individual human being, is conceived on the model of an object of natural or social scientific inquiry. Such action is taken to be a distinctive phenomenon, a distinctive type of event.

2. The inquirer, the one who is raising the question that the theory is to answer, is taken to be related to this phenomenon as a scientist to an object of scientific inquiry. It is from this standpoint that the inquirer regards the explanandum as needing explanation.

3. The philosophical significance of such action is that it is a type of event to which we attach a distinctive normative status. Actions are events that somehow "speak for" or are attributable to agents. Other events, like volcanic eruptions and allergic reactions, do not have this status. Moreover, we think of human actions as subject to normative standards. They can be rational or irrational, whereas volcanic eruptions and allergic reactions cannot.

4. The philosophical task, then, is to explain the what the phenomenon is, in such a way as to show us why it makes sense to attach this normative status to it.

The mechanist assumes his opponent shares this approach and that it is with reference to this methodological task that his opponent begs the question. But I will argue that Kant pursues a different approach. His critical philosophy is not modeled on natural and social scientific inquiry, and its aim is not to explain what happens when someone acts.[2] Rather, its aim is to show us what we are doing insofar as we are acting.

[2] Granted, Kant often describes his critical method as "scientific," but in those cases, I take it that he is using the term in a much broader sense.

IV Kant's critical method

Kant's theory of agency just is his theory of practical reason. His theory of practical reason, in turn, is part of his critique of reason as a whole. So let me step back and say something about how Kant frames this larger inquiry. The critique of reason is not conducted from the standpoint of a natural scientist. It cannot be, since one of its central aims is to map out the scope and limits of reason in its natural scientific employment. As Kant writes in the Introduction to the *Critique of Pure Reason*, "our object is not the nature of things, which is inexhaustible, but the understanding, which judges of the nature of things" (Kant 1998: 133/A12–13). Critique investigates our own reflective activity in each of its forms, for example, logical, mathematical, natural scientific, ethical, aesthetic. What is most important for our purposes is this: the theory investigates our own reflective activity from the standpoint of one who is already engaged in, and thus tacitly committed to, the reflective activity it investigates. The question that arises from this standpoint is not, "What happens when someone reasons?" Rather, it is, "What am I doing, insofar as I am reasoning?"

The sense of "What am I doing?" here is exactly the sense in which you might ask, "What am I *doing*?" upon discovering that you have drifted off track. It arises out of disorientation. Here is an example. Suppose you have been a journalist for the past ten years. Increasingly, commercial and political pressures have been influencing your profession, and you have grown disillusioned. "What am I *doing*?" you ask. By this you mean, "What am I doing, insofar as I am engaging in journalism?" Possible answers are: "Am I entertaining an audience? Am I increasing profits? Am I promoting a political agenda?" In asking this question, you are trying to distinguish journalism from closely related activities. But your interest is not simply taxonomic. It is different from, say, a geologist's interest in distinguishing one type of mineral from another or a medical researcher's interest in distinguishing one type of illness from another. What you are interested in is vindicating a form of activity to which you are already tacitly committed. You are asking what journalism is, because you tacitly take it to be a form of achievement, something worth doing, and you are interested in showing yourself how journalism could be the worthwhile undertaking you already taking it to be. You are asking, "What am I doing?" in the sense of, "What must I be doing if I am to be engaging in journalism, the activity I value, rather than entertainment or profiteering or propaganda?"

What you are asking for is something I will call an "ideal conception" of journalism. An ideal conception of an activity is a description under which you value the activity in

which you are engaged, a description under which you see it as worth undertaking.[1] We need ideal conceptions of what we do because, and insofar as, we self-consciously decide to do anything. For in order to decide to do anything, you have to conceive of *what it is* you are deciding to do. And in that moment, you have to conceive of *that* – what you are deciding to do – in terms that reveal its value to you, its choiceworthiness.[2] You need a conception that is at once descriptive and evaluative. This is not a sign of confusion. The idea that description and evaluation are fundamentally separate activities has great importance when the concept in question refers to an object that is related to us as a phenomenon to be explained scientifically. This is because the job of scientific description is to represent objects as part of an order of nature, and there are no values built into that order. But when the concept in question picks out an object that is differently related to us, it does a different job. When the concept is that of "What I am deciding to do?" applied from the participant standpoint, its job is to represent an activity to me as something that is worth undertaking, given my situation and my commitments. So construed, the role played by this concept has both a descriptive and an evaluative aspect. If I were to conceive of my candidate actions in purely empirical terms, as objects of scientific inquiry, I would not be able to engage in decision-making about whether to undertake them. It is not possible to undertake an event.

Kant's critique raises this "What am I doing?" question with respect to reasoning in each of its employments.[3] And it raises this question in exactly the sense I described, namely, "What must I be doing, if I am to be engaging in the activity I value, rather than some degenerate version of it?" Consider Kant's account of natural scientific reasoning. What are we doing when we are making the judgment that A causes B? Kant argues that we are making a synthetic *a priori* judgment, and he tries to offer an ideal conception of what this activity involves, such that we can endorse and take responsibility for engaging in it. Indeed, Kant thinks it is important to take responsibility for this cognitive activity, because our very capacity to put thoughts together in this way renders us capable of failing to do so. The point of critique is to help us stay on track. Its role is "only negative, serving not for the amplification but only for the purification of our reason, and for keeping it free of errors ... by supplying the touchstone of the worth and worthlessness of all cognitions a priori" (Kant 1998: 133/A12/B26). "Through criticism alone," Kant writes, "can we sever the very root of materialism, fatalism, atheism, freethinking unbelief ... enthusiasm, superstition ... and idealism and skepticism" (Kant 1998: 119/Bxxxiv).

[1] An "ideal conception," as I am using this concept, functions in the same way as a "practical identity" (Korsgaard 1996: 101), except that it describes an activity rather than a person. A practical identity is "a description under which you value yourself," whereas an ideal conception is a description under which you value what you are doing.

[2] Kant's notion of a "maxim," as I interpret it, plays just this role.

[3] As John Rawls notes, 'For Kant, pure reason ... is the faculty of orientation' (Rawls and Herman 2000, 263/Lecture IV, sec. 4). Rawls credits Susan Nieman with this conception.

Critique plays this negative role by working out an ideal conception of the activity in terms of its constitutive standards and aims.[4] This involves identifying constitutive concepts and principles that are essential to guiding participants engaged in the activity. To return to our previous example, in order to engage in journalism at all, you must gather information from sources. These concepts – "gather," "information," and "sources" – are essential to the structure of the activity. They describe a specific move you have to make, in relation to certain roles that have to be played, if journalistic achievements are to be possible through your efforts. With regard to this description, further questions arise. What sort of information is relevant to journalistic achievement? Which activities count as gathering it for that purpose, and which people or things count as sources? A deeper understanding of the constitutive aim of journalism, and of salient corrupting influences, will help answer those questions. Similarly, Kant maintains, there are concepts that are essential to the structure of natural scientific reasoning. A "category of the understanding" is one of them. It marks a specific role that has to be played if achievements in natural scientific reasoning are to be possible through our efforts. And with regard to this concept, there are further questions. Which concepts count as categories of the understanding? What is it to "apply" such a category to an object? A deeper understanding of the constitutive aim of natural scientific reasoning, and of salient corrupting influences, will help answer that question.

This is the sense in which Kant gives us a theory of the nature of theoretical cognition of objects. He is not doing cognitive science, if the aim of cognitive science is to explain what is happening when we are engaged in this kind of thinking. Rather, he is giving us an ideal conception what we are doing – indeed, what we *must* be doing – insofar as we are engaged in this form of thinking. This involves giving a constitutive anatomy of this activity, regarded as an undertaking and as a form of achievement.

[4] The critical method is thus inherently "constitutivist," and Christine Korsgaard's "constitutivist" reading of Kant simply follows from this method (Korsgaard 2009). David Velleman's constitutivism is importantly different. Velleman does not explicitly reject the question, "What happens when someone acts?", and yet I believe he ends up answering the question, "What am I doing insofar I am acting?" I believe there is a tension between these aspects of his view.

V Kant's theory of agency

How does this method shape Kant's theory of agency? From what I have said so far, it should be clear that Kant conceives of all forms of reasoning as involving agency in a broad sense. Reasoning is reflective activity, and reflective activity comes in various forms. Kant's theoretical philosophy addresses us as participants in theoretical reasoning, the aim of which is cognition of objects. His practical philosophy addresses us as participants in practical reasoning, the aim of which is determination of the will. How, then, does he conceive of the will? Recall the passage I cited at the outset: "Everything in nature works according to laws. Only a rational being has the capacity to act *in accordance with the representation* of laws, that is, in accordance with principles, or has a *will*" (Kant 1996a: 66/4:412). How should we read this statement? Is Kant describing an ontologically real entity, one that could play a role in an explanation of the events we call the "actions" of rational beings?

No. It is in keeping with the method of critique, as I have described it, that Kant is answering the question, "What is the nature of the will?" in the context of answering the more fundamental question, "What am I doing when I am engaging in practically reflective activity?" If this is right, then his description of the will should be read as part of an ideal conception of practically reflective activity. Kant is telling us what we are undertaking, and what form of achievement we are valuing, insofar as we are reasoning practically. The reflective activity in which we are engaged, and to which we are tacitly committed, is that of acting in accordance with our representation of principles. This is not a circular characterization. The claim is not that the reflective activity in which we are engaged is reflective activity. Rather, the claim is that the reflective activity is acting, determining ourselves in accordance with principles, in contrast with knowing, representing given objects. Agency in the practical sense is a species of the larger genus, reflective activity, a genus that includes various different forms of reasoning.

The theory proceeds by further anatomizing this activity in terms of its constitutive standards and aims (Kant 1996a).[1] In a nutshell, the aim is self-determination, and the standard is that the principle in accordance with which we determine ourselves is unconditioned by any prior need or interest. As Kant sees it, we construct our principles of action, but we can do so in a way that succeeds or fails to realize our capacity for self-determination. We do it in the latter way, a way Kant calls "heteronomous,"

[1] See especially the *Groundwork of the Metaphysics of Morals* and the *Critique of Practical Reason*, both contained in Kant (1996a).

when we allow our inclinations to pressure us into constructing principles simply to satisfy them. We do it in the former way, "autonomously," when we construct principles independent of that pressure. What must we be doing, insofar as we are constructing principles independent of that pressure? Kant argues we must be constructing principles in accordance with a certain formal constraint, a constraint that requires us to choose our principles first and foremost as principles and only secondarily as instruments to serve prior needs and interests. We are fully self determined insofar as we choose our actions as principled ways of meeting our needs and interests rather than simply as effective ways of meeting our needs and interests. The content of this ultimate, formal constraint, Kant tells us, is given by the various formulations of the categorical imperative.

The role of critique here is still negative, but the threat is not directly parallel to the threat we faced as theoretical reasoners. There the temptation was to allow pure reason to overstep its proper bounds by constructing empty thoughts and presenting them as if they had objective reality. Here the temptation is to allow empirical practical reason to usurp the role of pure practical reason, by constructing inclination-driven principles and conforming to them as if they were chosen autonomously (Kant 1996a: 148/5:15–16). In both cases, however, the philosopher's task is to correct and "purify" our thinking about what we are up to, separating out necessary from contingent elements of the activity, and showing us what conditions have to fulfilled in order for it to be possible as a form of achievement.

VI Why Kant thinks his method tells us what agency is

To sum up the main contrast: The mechanist's theory of agency is addressed to the inquirer as an observer of phenomena. Human action is conceived as a type of event. What is special or puzzling about this event is that it is attributable to an agent. The account tries to show what happens when someone acts such that the action-event counts as attributable to an agent. It does so by trying to identify a mechanism of causal (or functional) determination that has the status of being internal rather than external to the agent, in the sense that it speaks for the agent instead of simply moving her around.

Kant's theory of agency is addressed to the inquirer as one who is engaged in a form of activity, namely practical reasoning. Practical reasoning is conceived as one distinctive form reflective activity among others. What is special or puzzling about this form of activity is that it aims at self-determination. The account tries to show us what we must be doing, such that we succeed in determining ourselves. It does so by trying to identify a principle of choice that has the status of being internal rather than external to us, in the sense that it represents us rather than any particular need or interest in us.

Notice that the sense of "determination" is different in each approach. "Determination" in the context of the mechanist's approach refers to determination of an event by a law that explains the workings of the mechanism. "Determination" in the context of Kant's approach refers to determination of a choice according to a rule to which an agent self-consciously conforms. The mechanist is likely to object that insofar as Kant helps himself to this latter notion of "determination," he is still begging the question.

But that assumes the Kant shares the mechanist's way of framing the question. If Kant is not trying to explain what happens when someone acts, then he does not need to explain what happens when someone determines his choice according to a rule. Still, the mechanist may respond that indeed, Kant is pursuing a coherent approach to *a* question, but it is not *the* question about what individual human agency *is*. That question, the mechanist will claim, is about how human agency fits into the world. The point of the theory, he will argue, is to give us a representation of the world that shows how our agency is part of it.

This is where Kant's theory of transcendental freedom becomes relevant (Kant 1998: 484/A444/B472; Kant 1996a: 215–225/5:93–106; Allison 1990). Here I can only present that theory in bare outline, and I am admittedly presenting only one of several possible

interpretations of the texts. In claiming that the point of a theory of agency is to show how agency fits into the world, the mechanist is making an assumption about the role of the concept of agency. He is assuming that the concept of agency plays the role of an empirical concept, in that we use it to pick out an object in the order of nature. The spooky libertarian shares a version of this assumption with regard to the concept of freedom. He assumes we use the concept of freedom to pick out a non-empirical object, a contra-causal power which, though inexplicable, explains certain events in the order of nature. But Kant's critique of theoretical reasoning leads him to conclude that the concept of freedom cannot be used in either of these ways. Although reflection naturally leads us to construct a concept of freedom, we cannot apply it to objects in accordance with the rules of theoretical cognition. It is an empty idea, for explanatory purposes. Now as I am understanding Kant, this is very different from saying that freedom is a spooky power. I take it Kant is denying that the concept of freedom refers to any power, spooky or otherwise, that would play any role in explaining events, including our own behavior.

Nevertheless, Kant maintains that it is possible and fruitful to give freedom the status of a practical postulate (Kant 1996a: 246–7/5:132–4). What does this mean? Some might think that what Kant has in mind is this: when we deliberate, we are allowed to assume that we have a contra-causal power to determine events. Freedom is a necessary fiction, one we are allowed to indulged to serve a practical need.

But in that interpretation, we would be employing the concept of freedom as if it picked out a phenomenon that does explanatory work, even though we understand that this cannot be the case. We would be using a theoretical concept in bad faith. This would simply be self-deception.[1] I take it that Kant's point is different. It is that the freedom we are allowed to postulate for practical purposes is itself a practical concept. This concept does not even purport to pick out a phenomenon. Instead it specifies a form of activity, and a form of achievement, in which we are already engaged and to which we are already committed. When we conceive of our freedom this way, we are led to an ideal conception of what it is to determine ourselves. We determine ourselves by acting on the categorical imperative. Thus, the move from the theoretical to the practical sense of "determination" is justified by Kant's theory of the status of the concept of freedom. If the mechanist denies that freedom has this status, then he is making a substantive philosophical claim, for which he has to argue.

While Kant does not feel he needs to show how the phenomenon of human agency fits into the world, he does feel he needs to show how the concept of human agency fits into our lives. Indeed, he takes it as a task of his theory of reason as a whole to show how the different forms of reflective activity, along with their constitutive concepts and principles, fit together. The sort of unity needed for this is not the unity of a representation of objects of experience. It is the unity of a form of life, one in which the reflective subject is able to engage in each form of reflective activity without

[1] Kant rejects this sort of view in 1996a, 221/5:101.

undermining her own engagement in any other. If, in order to engage in practical reasoning, we had to believe as a necessary fiction that we have a spooky power, this would undermine our engagement in theoretical reasoning. Theoretical reasoning rules out the possibility of an event that does not have an efficient cause governed by natural laws. But theoretical reasoning cannot rule out the possibility that we can realize freedom as a reflective achievement. We can engage in practical reasoning, and in so doing strive to realize our autonomy, without committing ourselves to any judgment that natural science rules out (Bok 1998).

VII Conclusion

I have argued that a perennial and relatively intractable disagreement about how to characterize motivation can be rooted in a disagreement about why we need a theory of human agency. Kant's approach to the theory of agency becomes less mysterious when interpreted in this context. What may look like circularity is actually Kant's commitment to a method of critique, a mode of theorizing that is addressed to us as participants in reasoning rather than as observers of the phenomenon of reasoning. The critical aim is not to explain what happens when someone acts, but to show us what we are doing insofar as we are acting. Such a theory gives us an ideal conception of what we are doing when we are engaged in, and tacitly committed to, determining ourselves.

The broader methodological lesson is this. We should pause before assuming that the role of philosophical concepts, like that of empirical concepts, is to pick out phenomena. By the same token, we should pause before assuming that philosophical theories, like scientific theories, ought to explain phenomena. Science is one form of reflective activity among others. Kant's critical method expresses the view that philosophy is autonomous when it specifies the role of scientific thinking in our lives, rather than the other way around.

Further reading

Allison, Henry E. (2011). Kant's groundwork for the metaphysics of morals: A commentary. *Oxford: Oxford University Press.*

Herman, B. (1993). *The practice of moral judgment.* Cambridge: Harvard University Press.

Hill, T.E. (2002). *Human welfare and moral worth: Kantian perspectives.* Oxford: Oxford University Press.

Korsgaard, C.M. (1996). *Creating the kingdom of ends.* Cambridge: Cambridge University Press.

Reath, A. (2006). *Agency and autonomy in Kant's moral theory.* Oxford: Oxford University Press on Demand.

Reath, A. and Timmerman, J. (2010). Kant's 'critique of practical reason': A critical guide (Cambridge critical guides). *Cambridge: Cambridge University Press.*

Wood, A.W. (2007). *Kantian ethics.* Cambridge: Cambridge University Press.

References

Allison, H.E. (1990). *Kant's theory of freedom*. Cambridge: Cambridge University Press.

Bok, H. (1998). *Freedom and responsibility*. Princeton: Princeton University Press.

Bratman, M.E. (1987). *Intention, plans, and practical reason*. Cambridge: Harvard University Press.

Bratman, M.E. (2007). *Structures of agency: Essays*. Oxford: Oxford University Press.

Bratman, M.E. (2014). Rational and social agency: Reflections and replies. In: M. Vargas and G. Yaffe eds., *Rational and social agency: The philosophy of Michael Bratman*. Oxford: Oxford University Press. pp. 294–343.

Campbell, C.A. (1951). Is 'freewill' a pseudo-problem? *Mind*, 60 (240), pp. 441–465.

Chisholm, R. (1964). *Human freedom and the self*. Kansas City: University of Kansas.

Davidson, D. (1963). Actions, reasons, and causes. *The Journal of Philosophy*, 60 (23), pp. 685–700.

Kant, I. (1996a). *Practical philosophy*. Cambridge: Cambridge University Press.

Kant, I. (1996b). *Religion and rational theology*. Cambridge: Cambridge University Press.

Kant, I., Guyer, P. and Wood, A.W. (1998). *Critique of pure reason (CPR)*. Trans. and Eds. Paul Guyer and Allen W. Wood. Cambridge: Cambridge University Press.

Korsgaard, C.M. (1996). *The sources of normativity*. Cambridge: Cambridge University Press.

Korsgaard, C.M. (2009). *Self-constitution: Agency, identity, and integrity*. Oxford: Oxford University Press.

Korsgaard, C.M. (2014). The normative constitution of agency. In: M. Vargas and G. Yaffe eds., *Rational and social agency: The philosophy of Michael Bratman*. Oxford: Oxford University Press. pp. 190–215.

Leibniz, G.W. and Clarke, S. (2000). *Leibniz and Clarke: Correspondence*, ed. A. Roger. Indianapolis: Hackett Publishing.

Rawls, J. and Herman, B. (2000). *Lectures on the history of moral philosophy*. Cambridge: Harvard University Press.

Velleman, J.D. (1992). What happens when someone acts? *Mind*, 101 (403), pp. 461–481.

11. Anscombe on acting for reasons

Keshav Singh

Introduction(1)

This chapter discusses some of G.E.M. Anscombe's contributions to the philosophy of practical reason. In particular, it focuses on her account of what it is to act for reasons. The major work in developing this account takes place in Anscombe's incredibly rich monograph *Intention*, though I will also draw on insights contained in her later writings.

Anscombe's *Intention* is widely considered a foundational text in contemporary philosophy of action. Frederick Stoutland writes in his introduction to a volume of essays on *Intention* that it "definitely established philosophy of action as a distinctive field" (2011, p. 5). Anscombe's work has also had some influence in the philosophy of practical reason. However, it has not received nearly as much uptake there as it has in the philosophy of action. And even when Anscombe is cited in work on practical reason, it is often only in passing.[1] And on the subject of acting for reasons in particular, the depth of Anscombe's contributions are often overlooked by philosophers of practical reason.[2] As I will discuss later, some of this may be due to the mistaken view that her contributions have been largely superseded by those of Donald Davidson and his followers.

I will not attempt to address in a single short chapter all of the rich contributions made by Anscombe (in *Intention* and elsewhere) to the philosophy of practical reason. Instead, I will focus on giving an opinionated introduction to what her work says about acting for reasons and how it can inform current theorizing on the matter. As I will show, Anscombe's views cut deeply against much of the current orthodoxy on acting for reasons and are worth taking more seriously in the philosophy of practical reason.

[1] For example, Dancy's (2000) *Practical Reality* cites Anscombe only in passing, despite the fact that Anscombe's views are highly congenial to Dancy's rejection of Davidson's view of acting for reasons. For further evidence of such a trend, see notes 17 and 18. Notable exceptions, however, are Vogler (2001) and Wiland (2012, ch. 7). Wiland's chapter, in particular, is one of the few dedicated to Anscombe on reasons.

[2] This is not to say that philosophers have completely overlooked Anscombe on acting for reasons. Philosophers of action – Anscombeans especially – often discuss her views thereof in the context of her overall theory of action (see, for instance, Thompson 2008, Wiseman 2016 and Ford 2017). My concern is that work done by philosophers of practical reason that is *in the first instance* about acting for reasons has overlooked Anscombe. This is what I will attempt to begin to remedy, and for this reason, I will not focus on reconstructions of Anscombe's overall theory of action by philosophers of action. Thanks to Jack Samuel for suggesting I clarify this.

(1) I owe thanks to Ruth Chang, Jack Samuel, and Eric Wiland for their helpful comments on an earlier draft of this chapter.

For Anscombe, the question 'What is it to act for reasons?' is intimately related to the question 'What is it to act intentionally?' I will begin by discussing that relationship. I will then further explicate Anscombe's view by discussing her rejection of two related views about acting for reasons: causalism (the view that reasons are a kind of efficient cause of actions) and psychologism (the view that reasons are mental states like desires and beliefs). In the process, I will try to show that Anscombe's rejection of these two views does not leave us with mystery but rather sheds light on an interesting, heterodox account of acting for reasons.

1 Intentional action and acting for reasons

Anscombe argues in *Intention* that what distinguishes actions that are intentional from those that are not is that intentional actions are those "to which a certain sense of the question 'Why?' is given application." This is the sense in which "the answer, if positive, gives a reason for acting" (1957, p. 9).[1] We can already see here the intimate connection between acting intentionally and acting for reasons. It is even tempting to assume, based on this remark, that acting intentionally and acting for reasons are just the same thing for Anscombe. But this would be a faulty assumption, for the applicability of the special sense of 'Why?' is broader than the set of cases where the agent acts for reasons.[2]

This is because Anscombe's question 'Why?' applies in the relevant sense even in cases where the answer to the question is 'no particular reason.' As she says, "the question is not refused application because the answer to it says that there is *no* reason, any more than the question how much money I have in my pocket is refused application by the answer 'None'" (p. 25). In her view, acting intentionally is necessary but not sufficient for acting for a reason. So, acting intentionally cannot be the same thing as acting for a reason.[3]

In what cases, then, does the question 'Why?' *fail* to apply in the relevant sense? Anscombe discusses three circumstances where it fails to apply. Since these are circumstances in which one fails to act intentionally, they are also circumstances in which one fails to act for a reason. The first circumstance is when the answer is to the effect of 'I didn't know I was doing that.' To respond (sincerely) as such is to refuse application to the special sense of the question 'Why?'. For example, imagine I am spraying grass

[1] Throughout this chapter, all quotations from Anscombe are from *Intention* unless otherwise noted.

[2] If this is right, it raises questions about the relation between acting for no reason and *arational* action. One natural view is that if an action is done for no reason, it is thereby arational. But for Anscombe, actions done for no reason are still intentional. Moreover, insofar as the question 'Why?' is granted application, such actions are still in some sense intelligible. This raises the possibility that actions done for no reason are not thereby arational. While I cannot discuss this possibility at length here, it merits further exploration. Thanks for Ruth Chang for raising this possibility.

[3] This may not be an uncontroversial interpretation (though, to my knowledge, there is not a lot of work that directly addresses this question). Some of the remarks in Thompson (2008) suggest that, in his interpretation of Anscombe, acting for a reason and acting intentionally are coextensive. Thanks to Jack Samuel for pointing this out.

killer on my lawn, thinking it is weed killer. You, knowing that it is grass killer, ask me "Why are you spraying grass killer on your lawn?" and I respond, "This is *grass* killer!?". My response refuses application to the question 'Why?' because it indicates that I was not aware I was spraying grass killer on my lawn. Correspondingly, I was not intentionally spraying grass killer on my lawn.

Importantly, one might know what one is doing under some descriptions but not others. If you instead ask me "why are you spraying your lawn?" I might respond, "to kill these weeds over here." In this case, because I am aware that I am spraying my lawn, the question has application, and my response specifies a reason for my action. Correspondingly, then, I might act intentionally under some descriptions and unintentionally under others: while I am spraying my lawn intentionally, I am not spraying grass killer on my lawn intentionally.[4]

Although acting intentionally and acting for reasons are not the same, whether we have acted for some particular reason also seems to depend on the description of the action. It is unobjectionable to say that I sprayed my lawn for the reason that doing so would kill the weeds over there, but would strike us as quite odd to say that I sprayed grass killer on my lawn for the reason that doing so would kill the weeds over there. The latter statement makes me out to be wittingly instrumentally irrational when in reality I am just unaware of what I am doing. In the Anscombean view, the reasons for which we act stand in relation to our actions only under those descriptions under which they are intentional.

The second circumstance in which the question 'Why?' is refused application is when, despite the agent's awareness of what she is doing, her action is involuntary. As Anscombe notes, it is difficult to further cash out the notion of the involuntary without presupposing an account of intention. However, she writes, there is "a class of the things that fall under the concept 'involuntary' which it is possible to introduce without begging any questions or assuming that we understand notions of the very type I am professing to investigate" (p. 13). This is the class of bodily movements that are described in physical terms, but that one nevertheless knows without observation. Anscombe's own example is "[t]he odd sort of jerk or jump that one's whole body sometimes gives when one is falling asleep."This is an important category to mention because, for Anscombe, intentional actions are known without observation. However, some actions that are known without observation fail to be intentional because, being involuntary, the question 'Why?' fails to apply to them.

The role of the Anscombean thesis that when we act intentionally, we know what we are doing without observation is a fraught issue in Anscombe scholarship and is not the focus of this chapter. Nevertheless, Anscombe's account of non-observational knowledge ends up shedding important light on what it is to act for reasons. I will return to this issue in §3 and bracket non-observational knowledge for now.

[4] This raises the question of whether this is really the very same action under different descriptions or distinct actions. For an illuminating discussion of Anscombe's views on this matter, see Annas (1976).

The third circumstance in which the question 'Why?' fails to apply is when "the answer is evidence or states a cause, including a mental cause" (p. 24). Imagine I say I will go for a run, and you ask me "Why?"If I respond with evidence that I will run (i.e. what I take to be a reason to think I will run), then I am making a prediction instead of expressing an intention. By contrast, if I answer by giving what I take to be a reason *to run*, I am expressing an intention to run. One difficulty in interpreting Anscombe's remarks here is the fact that she does not explicitly distinguish what we now call *normative* reasons (reasons that count in favor of doing or thinking something) from *motivating* reasons (reasons for which we act, believe, and so on). If we assume that when she writes about acting *for a reason*, the reasons she picks out best fit into the category of *motivating* reasons, then one interpretation of these remarks is that an answer to the question 'Why?' that specifies the agent's motivating reason must give what is, from her perspective, a normative reason to act – that is, an answer that shows there to be something worthwhile in performing the action, rather than one that simply counts in favor of thinking it will happen.[5] I will return to these issues in §2 when I discuss what makes something the reason for which one acts.

Anscombe writes that the question 'Why?' fails to apply not only when the answer cites evidence but also when it cites a cause. Since the question 'Why?' does have application when the answer cites a reason, this suggests that, for Anscombe, reasons and causes are altogether different things.[6] Of course, everyone theorizing about the reasons for which we act makes a distinction between reasons and *mere* causes – that is, no one thinks that just any cause of an action counts as a (motivating) reason for that action. Anscombe, however, is plausibly interpreted as arguing for something much stronger: that reasons are not causes at all. If this is right, it is one of the most distinctive implications her theory has about the nature of acting for reasons. So, it is to this issue that I now turn.

See also Anscombe's own essay "Under A Description" (1979).

[5] Given that Anscombe doesn't use the terminology of motivating and normative reasons, this is not an uncontroversial interpretation. But I think it is a plausible one that helps to make sense of her views about how the reasons for which we act relate to our actions.

[6] Of course, this doesn't entail that that which is a reason can never also be what happens to cause an action. But even if it did happen to cause an action, for Anscombe, this would have nothing to do with what makes it a reason. Thanks to Eric Wiland for suggesting I clarify this.

2 Reasons and causes

In the contemporary literature on both intentional action and acting for reasons, the standard story has come to be a causal story. This is, as Stoutland writes, largely owing to the influence of Davidson's work on these matters:

> The majority of contemporary philosophers accept what has been called the "standard story" of action. That there is a standard story is largely due to Davidson, and he is usually said to accept it. It comes in different versions, however, many of which diverge from Davidson's own view to a greater or lesser extent. What unifies them is commitment to the claim that an agent's acting consists of those events that are her bodily movements caused (in the right way) by the beliefs and desires that are her reasons for acting.
>
> (2011, p. 12)

Of course, Davidson's own account of the reasons for which we act was a response to what he saw as a mistaken attack by Anscombe and others on the "commonsense" position that the relation between our actions and the reasons for which we act is a "species of ordinary causal explanation" (1963, p. 685). Davidson thought that with some refinement, he could accommodate Anscombe's insights without departing from the causalist picture of acting for reasons.

Despite Davidson's ambitions, Anscombe's own views paint a picture of acting for reasons that is radically incompatible with causalism. As Rosalind Hursthouse writes:

> [Anscombe's picture] has been obscured by the prominence of the current causal theory according to which intentions are antecedent events which explain intentional action. Indeed I have found in discussion that some people, perhaps misled by Davidson's many acknowledgments of *Intention's* insights, assume that all the book contains of value is some gropings towards the theory he has been developing since *Actions, Reasons and Causes*. But nothing could be further from the truth. *Intention* stands as an account of intentional action totally opposed to any causal account and not in need of radical development or improvement.
>
> (2000, p. 83)

In light of Davidsonian dogma, Anscombe's views about acting for reasons have become deeply heterodox. And my sense is that Hursthouse was right to claim that these views have been underappreciated and underexplored by those who dismiss out of hand the possibility of a non-causal picture of acting for reasons.

For Anscombe, to act for a reason is not for one's action to be caused (in the right way) by one's beliefs and desires. In fact, acting for reasons is not a matter of mental causation at all. The reasons for which we act, in her view, are neither mental states nor causes. In later work, she chalks the Davidsonian view up to a "failure of percipience" owing to "the standard approach by which we first distinguish between 'action' and what merely happens, and then specify that we are talking about 'actions'" (2006b, p. 111). To be fair, causalism's status as dogma is due not just to the shortcomings of Davidsonians but also the apparent opacity of the non-causalist picture. After all, Anscombe is reluctant to give a definitive statement of what she thinks it is to act for reasons. Nevertheless, her view can be reconstructed in a way that is not only of exegetical interest but demonstrates the seriousness of her brand of non-causalism.[1]

Most illuminating of her views on reasons versus causes, perhaps, is Anscombe's discussion of 'motives' (which she seems to take as interchangeable with 'reasons for action,' in the sense of motivating reasons) and the contrast she draws with mental causes. Anscombe identifies mental causes as "what someone would describe if he were asked the specific question: what produced this action or thought or feeling on your part: what did you see or hear or feel, or what ideas or images cropped up in your mind, and led up to it?" (pp. 17–18). In the causalist view, a motive just is a particular kind of mental cause – a mental state that causes our actions in right way to be potentially rationalizing. Anscombe disagrees with this view:

> Motives may explain actions to us; but that is not to say that they 'determine', in the sense of causing, actions. We do say: 'His love of truth caused him to ...' and similar things, and no doubt such expressions help us to think that a motive must be what produces or brings about a choice. But this means rather 'He did this in that he loved the truth'; it interprets his action.
>
> (p. 19)

Motives, then, do not cause action. Rather, they interpret action; they make sense of it.[2] They need not involve what went on in my mind prior the action and issued

[1] Furthermore, it demonstrates that non-causalists about reasons for action themselves should take Anscombe more seriously than they do, not least because she is in some sense on their side. See note 18 for some evidence that even non-causalists have not paid enough attention to Anscombe's view of acting for reasons, despite its congeniality.

[2] Importantly, however, this should not suggest a picture where interpretation is something we 'add' to a prior distinct existence. For Anscombe, that would raise the same problems as causalism. The interpretation is not something we add to an action but rather part and parcel of it. Thanks to Eric

in the action. Instead, they are features of the action itself through which one makes sense of it, as indicated by the answers one would give when asked the question 'Why?'

Anscombe's discussion of three kinds of motives sheds further light on how motives relate to actions. The three kinds of motives are backward-looking motives, motives-in-general, and forward-looking motives. Backward-looking motives are things like revenge, gratitude, pity, and remorse. In the context of explaining such motives, Anscombe addresses the question "Why is it that in revenge and gratitude, pity and remorse, the past event (or present situation) is a reason for acting, not just a mental cause?" (p. 21). Of course, the causalist would claim that such motives are the reason for which we act in virtue of being mental causes of a certain sort. But Anscombe wants to show how they are reasons *not* in virtue of being mental causes of any sort. She contends that backward-looking motives are distinct from mental causes because they consist in the agent's conceiving of them "as something good or bad, and his own action as doing good or harm" (p. 22).

Here we finally get something close to a statement of a condition on something's being a (motivating) reason for action, because what Anscombe actually says is that an agent's answer to the question 'Why?' is a reason for acting if in treating it as a reason he conceives of it as described in the previous quote. This suggests that, though reasons are neither mental states nor causes, acting for reasons does involve the mental states of the agent.[3] Indeed, it would be difficult to see how an agent could act for reasons without her mental states being involved in some way. Nevertheless, some consideration does not become one's motivating reason by being the right kind of mental state, that causes in the right way, one's action. Instead, for Anscombe, some consideration becomes one's motivating reason by being represented by the agent as standing in some kind of relation to the action that makes sense of the action from her perspective. To answer the question 'Why?' by providing a motive-in-general, Anscombe writes, "is to say something like 'See the action in this light'" (p. 21). Forward-looking motives, of course, simply specify future ends to which the action would be a means. All three kinds of motives *make sense* of the action from the agent's perspective.

Anscombe's discussion of motives is helpful partly because it shows us what kind of answers to the question 'Why?' reasons are supposed to be. While causalists also think of reasons as answers to the question 'Why?', they conceive of reasons as causal explainers of action, so they consider the question "Why?' a request for a particular kind of causal explanation. Anscombe, by contrast, conceives of reasons as that which, from the agent's own perspective, make sense of what she is doing. For Anscombe, the question 'Why?' is a request not for a causal explanation at all but a *sui generis* kind of explanation (what we might call an interpretive explanation). Of course, the causalists think we can give an interpretive explanation just by giving the right kind of causal explanation. As Davidson writes,

---

Wiland for suggesting I clarify this.

[3] It is worth flagging that Anscombe has qualms about the language of mental *states* in this context.

> A reason rationalizes an action only if it leads us to see something the agent saw, or thought he saw, in his action – some feature, consequence, or aspect of the action the agent wanted, desired, prized, held dear, thought dutiful, beneficial, obligatory, or agreeable.
>
> (1963 p. 685)

One way of casting the disagreement between Anscombe and the causalists, then, is that Anscombe is an anti-reductionist about interpretive explanations of action. By contrast, causalists are reductionists; they think interpretive explanations can be reduced to a species of ordinary causal explanation.

Here is what we have so far. In Anscombe's view, a reason for action is an answer to the question 'Why?' that explains the agent's action by showing what made sense of the action from her perspective, whether that be a means-end relation, an interpretation of the action in some larger light, or some backwards-looking motive like revenge. Importantly, reasons *do not* explain actions by causing them. For Anscombe, reasons and causes stand in stark opposition. In the next section, I will try to say a little bit more about why she takes her view to be unassimilable to the causalist picture, as well as what she takes the non-causal relation between reasons and actions to be.

But for my purposes here, it won't be problematic to stick to that language while noting her reservations.

3 Reasons as constituents of action

Anscombe is clear that in her view, neither intentions nor reasons are causes of actions. Part of why she is so insistent on this is that for either to be causes, they would have to be separable from the actions they caused. But this, in her view, leads to all sorts of problems. This is why she criticizes Davidson for conceiving of intentional actions as events to which we affix certain additional, extrinsic features, like their being caused by certain mental states. She thinks this aspect of the causal view subjects it to at least one fatal problem: the problem of deviant causal chains.

The problem of deviant causal chains is a challenge to the causalist's ability to give an account of what it is for the relevant mental states to cause an action 'in the right way.' Consider Davidson's (1973) famous example of a climber who wants to rid himself of the weight and danger of holding another man on a rope and believes that loosening his grip on the rope will accomplish this. This belief-desire pair so unnerves the climber that he inadvertently loosens his grip on the rope. In this case, the belief-desire pair causes the climber's action but does not seem to cause it the right way to be his reason for action. Here is what Anscombe has to say about such examples:

> Davidson indeed realized that even identity of description of *act done* with *act specified in the belief*, together with causality by the belief and desire, isn't enough to guarantee the act's being done *in pursuit of the end* and *on grounds of the belief*. He speaks of the possibility of 'wrong' or 'freak' causal connexions. I say that any recognizable causal connexions would be 'wrong', and that he can do no more than postulate a 'right' causal connexion in the happy security that none such can be found. If a causal connexion were found we could always still ask: 'But was the act done for the sake of the end and in view of the thing believed?'
>
> (2006b, p. 110)

So, clearly Anscombe is pessimistic about the possibility of identifying the 'right' causal connection, because in her view, *no* causal connection could establish that the agent acted for the sake of a particular end. And indeed, Davidson and his successors in the causalist tradition have struggled to find a satisfying solution to the problem of deviant causal chains, providing some inductive support for Anscombe's pessimism.[1]

[1] Even Davidson himself became a kind of defeatist about the problem (see Davidson 1973). Recently, some causalists have attempted to solve the problem by appealing to dispositions. For example,

Since Anscombe rejects the causal story, we need an alternative picture of how the agent's own understanding of why she is doing what she is doing makes it the case that she is acting for particular reasons. For the causalist, the agent has some mental states that at once cause her action and encapsulate what she takes to favor of performing it. Thus, the problem of deviant causal chains notwithstanding, causalists have a picture of how the agent's outlook on her action relates to the action itself. For her position to be plausible, Anscombe must have her own story about how these two things relate, and it must not fall prey to the very same problem that she takes to be fatal for the causalist.

Again, Anscombe is loath to offer a single, succinct statement that can be used to pin down her views on this matter. But her views can be reconstructed in a way that shows her non-causalism to be backed by a coherent metaphysics. Some of what she says in *Intention* may give the impression that she denies wholesale that any mental states or events could be relevant to a genuine interpretive explanation of an action, leaving it mysterious what actually renders the action intelligible from the agent's perspective. But this would be a misunderstanding of her view, as some of her remarks in later work clarify. In "The Causation of Action," she writes that "the teleology of conscious action is not to be explained as *efficient* causality by a condition, or state, of desire" [my emphasis] (2006a, p. 96). She then considers the objection that there must have been something in the agent's mind that "suffused" it with intentional-ness. She responds, "Was that then a separable mental experience which you want to say *caused* the action? ... in this conception a cause has to be thought of as a distinct thing, which is found to have this effect." The answer must be 'no,' because the mental state must be "intrinsic to an action when it is intentional; or rather, definable only by the description of the intentional action. But such is not a cause of the action" (pp. 96–97).

The foregoing makes it clear that what Anscombe objects to is not the involvement of the agent's mental life in any form in her action. That would be patently absurd. Instead, what she objects to is the reification of that mentality into a distinct existence that is then said to cause the action. And the problem of deviant causal chains provides principled grounds for her objection.[2] So, the first thing that has been clarified here is that Anscombe's rejection of the causal picture does not commit her to some occult view on which the agent somehow makes sense of her action without her mental life playing any role in explaining it.

Now we can finally ask: How, for Anscombe, does the agent's mental life contribute to her action? And how does the agent's understanding of why she does what she does contribute to the reasons for which she acts? The answers to these questions can be found in *Intention* itself. Here the role of non-observational knowledge in Anscombe's theory of intentional action becomes relevant. Quoting Aquinas, Anscombe writes that

see Wedgwood (2006), Hyman (2014), and Lord (2018).

[2] Anscombe also makes some interesting remarks about regresses created by the causal picture, which I don't have space to discuss here.

we have non-observational knowledge of what we are doing because such knowledge is "the cause of what it understands. This stands in contrast to observational knowledge, which is "derived from the objects known" (p. 87). The invocation of cause in the current context may seem strange, since we have just recounted Anscombe's hostility to the idea that causation by any mental states of the agent could be what separates intentional action from mere happenings.

The key here is that the term 'cause' in this context does not refer to the efficient causation we have been discussing so far. It refers instead to something like Aristotelian formal causation. As John Schwenkler puts it,

> at the core of Anscombe's account of action is the idea that practical thought is not an *efficient* cause that sets the visible parts of our body into motion, but the *formal* principle that *unifies* an action, or that in virtue of which certain physical happenings are constituted as parts of a person's intentional activity.
>
> (2015, p. 6)

So, for Anscombe, we have non-observational knowledge of what we are doing because the agent's own understanding of what she is doing is what constitutes it as her action. It is in this sense that non-observational knowledge of intentional action is the (formal) cause of what it understands.

The Aristotelian distinction between efficient and formal causation has as a rough analogue in contemporary theorizing the distinction between causation and constitution. As such, in what follows, it will be helpful to contrast causation with constitution. In Anscombe's view, our mental lives contribute to our actions very differently from how they contribute on the causalist picture. Our actions are not intentional in virtue of being caused by mental states like belief and desire. Rather, they are intentional in virtue of having a certain structure (of which the paradigm is a teleological structure) that is constituted by our very representation of it as having that structure. For example, imagine I wave my hand in the air in order to get your attention. What makes it the case that I wave my hand for the sake of this particular end? For the causalist, it is that I desired to get your attention and believed that waving my hand would do so, and this belief-desire pair causes me, in right way, to wave my hand. For Anscombe, it is simply that in waving my hand, I understood myself as getting your attention.

Now, in such a case, we can say that my reason for waving my hand was that it would get your attention. This makes it clear that for Anscombe, the reason for which one acts is not a distinct existence from the action itself. Instead, it is part of the structure of that action. Reasons, then, are related to our mental states in an important way – just not in the way causalists think they are. Causalists think they are the mental states that cause our actions. For Anscombe, they are the contents of mental states through which our actions are structured. In the paradigm case of a teleologically structured action, the agent constitutes a particular consideration as the

reason for which she acts by giving her action a structure where she understands it under a particular description (waving my hand) as a means to the end described by that very consideration (that it will get your attention).

This identification of means and ends with reasons and actions extends throughout series of multiple means and ends. For example, Anscombe's famous case where the man moves his arm, to operate the pump, to replenish the water supply, to poison the inhabitants. In each case, the end serves as the motivating reason for the action under the description of the means. And the final, non-instrumental motivating reason in the series is the final end: to poison the inhabitants. Of course, the finality of the end of poisoning the inhabits doesn't entail that the man does it for no reason. It is just that we must identify his reason for poisoning the inhabitants as something outside of the teleological structure – perhaps by providing some general or backwards-looking motive.[3]

Anscombe rejects the causalist picture partly on the basis of the problem of deviant causal chains. She takes this problem to arise because on the causalist picture, actions and the reasons for which we perform them are distinct existences and the fact that we perform an action for some particular reason is extrinsic to that action. In her view, by contrast, the reasons for which we perform an action are intrinsic to that action and not distinct existences. This is supposed to immunize her view from the problem of deviant causal chains, giving it a distinct advantage over the causalist view.[4] If reasons and actions are related constitutively, not causally, there is no mystery of how the reason and action relate in the right way to provide the relevant kind of explanation of the action in terms of the reason.[5]

[3] For more on this, see Stoutland (2011) on reasons that are internal to the teleological structure of an action versus those that are external to it.

[4] However, for an argument that non-causalists face an analogous problem, see Paul (2011) on 'deviant formal causation.'

[5] The idea that the reasons for which we act are related constitutively to our actions themselves evokes some prominent interpretations of Kant that also deny that reasons and actions are distinct existences. On this, see especially Korsgaard (2008, pp. 227–229), who writes that a reason for which one acts "is not a mental state that precedes the action and causes it," but is rather "embodied in the action itself." To my knowledge, this parallel between Kant and Anscombe has not been explored in depth. While it is unfortunately beyond the scope of this chapter, it certainly merits further exploration. Thanks to Ruth Chang for bringing it to my attention.

4 Anscombe's view in context

As we have seen, Anscombe's view of acting for reasons differs dramatically from what has become the Davidsonian orthodoxy. In Anscombe's view, the reasons for which we act are answers to the question 'Why?' in the special sense that calls not for the causes of the action but for the agent's own interpretation of what she is doing. Furthermore, Anscombe's discussion of motives makes clear that reasons for actions are not mental states but rather those considerations that make sense of the action from the agent's perspective.

As such, Anscombe rejects two commonly held views about motivating reasons. The first is a view about their ontology: the view that they are mental states, which has come to be called psychologism.[1] The second is a view about their relation to action: the view that they are causes of action, which I have been calling causalism. Following Davidson, almost everyone who accepts psychologism accepts causalism, though there is no inconsistency in accepting the former without the latter. And even those who reject psychologism often maintain causalism as the default position, perhaps in part because they cannot envision any plausible alternative to it.[2]

Relatedly, epistemologists writing about the epistemic basing relation (the relation between a belief and the reasons for which it is held) also tend to be causalists.[3] The dominance of causalism among both philosophers of action and epistemologists is naturally traced back to the Davidsonian thought that this is the only plausible way of understanding how reasons explain.[4] Part of the importance of reconstructing Anscombe's view, then, has been to show that this is false. Anscombe's non-causalism

[1] Aside from Davidson, the most prominent defender of psychologism is Smith (1994, 2003). In a similar vein, Turri (2009) defends psychologism about *believing* for reasons. For arguments against psychologism that are separate from Anscombe's, see Dancy (2000), Alvarez (2010, 2016), and Singh (2019). O'Brien (2015) also defends a non-psychologistic account of reasons explanations. Strikingly, however, though O'Brien engages substantially with Davidson, she does not even mention Anscombe.

[2] For dissent, however, see Dancy (2000) and Ginet (2002). For a defense of causalism from such dissent, see Davis (2005). As further evidence that Anscombe's contributions have recently been underexplored in comparison to Davidson's, all of the writings just mentioned engage substantially with Davidson but cite Anscombe only in passing.

Some accounts don't take a stand on whether the explanatory relation between reasons and actions is causal. Such accounts, as long as they reject psychologism, are in principle compatible with Anscombe's view. For examples, see Alvarez (2010, 2016) and Singh (2019), the latter of which is particularly congenial to Anscombe.

[3] See, for example, Wedgwood (2006) and Boghossian (2014).

[4] For example, Wedgwood cites Davidson as providing "[t]he principal argument for regarding the basing relation as a causal relation" (p. 661).

is a contender when it comes to reasons explanations just as much as when it comes to intentional action. But this is something even recent non-causalists (such as Dancy and Ginet) have tended to overlook. Anscombe's views thus have implications for current work not just on acting for reasons in particular but on related work on the epistemic basing relation and on the more general subject of reasons explanations.

There is much more to be said about how Anscombe's views on acting for reasons can inform our current theorizing; unfortunately, much of it is beyond the scope of this chapter. However, there is one issue I would like to discuss before concluding, and that is the issue of whether we can be ignorant of or mistaken about the reasons for which we act. Partly in light of recent work in psychology that purports to show our limited grasp of our own motivations, it has become a common view among philosophers that the reasons for which we act are often not transparent to us.[5] Indeed, some philosophers reject the idea that we have *any* privileged access to our motivating reasons.

Given her views about non-observational knowledge of what we are doing, Anscombe tends to be on the side of thinking that we have a very strong privileged access to facts about our actions, including the reasons for which we act. Anscombe's views, then, might be accused by current theorists of painting an unrealistic picture of human psychology and the transparency of motivation. In light of this, it is worth briefly examining Anscombe's views in light of the current consensus that our motivations are often opaque to us, even when we act for reasons (as opposed to mere causes).

Anscombe addresses this issue in a passage that is less commonly discussed but that seems to me to be of great interest:

> An answer of rather peculiar interest is: 'I don't know why I did it'. This can have a sense in which it does not mean that perhaps there is a causal explanation that one does not know. It goes with 'I found myself doing it,' 'I heard myself say ... ,' but is appropriate to actions in which some special reason seems to be demanded, and one has none... . I myself have never wished to use these words in this way, but that does not make me suppose them to be senseless. They are a curious intermediary case: the question 'Why?' has and yet has not application; it has application in the sense that it is admitted as an appropriate question; it lacks it in the sense that the answer is that there is no answer. I shall later be discussing the difference between the intentional and the voluntary; and once that distinction is made we shall be able to say: an action of this sort is voluntary, rather than intentional.
>
> (pp. 25–26)

[5] Much of this influence comes from Nisbett and Wilson (1977) and the following literature, which purports to show that people are prone to confabulating the reasons for which they act, and in such cases are unable to provide the 'true' explanations of their actions. For examples of such influence, see Nichols and Stich (2003), Gertler (2011), Carruthers (2013), and Cassam (2014). For criticism of the way Nisbett and Wilson's results are interpreted, see Sandis (2015).

I cannot undertake a discussion of Anscombe's distinction between the voluntary and the intentional here. But one thing is clear: in her view, an action fails to be intentional when the agent's answer to the question 'Why?' is 'I don't know why I did it.' For Anscombe, while this does not straightforwardly refuse application to the question 'Why?', it is not a genuine answer because it does not shed light on the agent's action from her own perspective. Given that the reasons for which one acts must be answers that shed such light on one's action, it seems that, in Anscombe's view, the agent cannot have acted for reasons in such cases.

This is compatible, as she says, with there being some causal explanation of his action that he does not know. However, she denies that in such cases, it is possible that there is "a reason, if only he knew it." Indeed, she denies that this is possible "even if psychoanalysis persuades him to accept something as his reason" (p. 26). This suggests that Anscombe's view radically diverges from current orthodoxy on the opacity of our reasons. The case of psychoanalysis is one that many current theorists would view as a paradigmatic case in which one might, through self-examination, discover the reasons for which one performed past actions. For Anscombe, such discovery is impossible. To act for a reason is to act in light of a consideration that makes sense of one's action from one's own perspective; this is impossible without having access to what that consideration is. Thus, we cannot be alienated from our reasons in the way that it is often thought we can.

The divergence between Anscombe and current orthodoxy on the transparency of our reasons is not unrelated to the divergence between them on causalism and psychologism. In the Davidsonian view, the reason is not only a distinct existence from the action but can be pulled apart from the agent's own perspective. Of course, Davidson grants that the agent see something good or worthwhile in the action. But this too can be divorced from her perspective, for in Davidson's view, this is just a matter of her beliefs and desires. So, we can be just as alienated from our reasons as we can from our beliefs and desires.[6] To see what an agent's reasons are, we need only discover which mental states caused her action, whether or not she knows what they are.[7] This is not possible for Anscombe.

One kind of case Anscombe does not explicitly discuss – one with which theorizing about transparency and alienation is concerned – is a case in which an agent cites what from her perspective is the reason for which she acted, but from a third-personal perspective, there is evidence that she is mistaken about why she acted. Most current theorists would want to hold that in such a case, the agent is indeed mistaken about the reasons for which the acted and that it is possible that what moved her was some unconscious motive rather than what she cited as her reason. It seems Anscombe must

[6] It's generally accepted that our desires can be non-transparent to us. If motivating reasons are desires or desire-belief complexes, then it follows that our motivating reasons can be non-transparent to us. For an example of this view, see Smith (1987).

[7] Indeed, given Davidson's interpretationism, it should be possible for others to know an agent's reasons for acting better than she herself does. See Davidson (1984) for more on his interpretationism.

deny the possibility of this case as well. For Anscombe, while there may be some mental cause of which the agent is unaware, this is irrelevant to the question of the reason for which she acts. Only the agent's own understanding of her action matters. Perhaps Anscombe would attribute the claim that there are cases of being mistaken about one's reasons to the erroneous assumption of causalism.[8]

Anscombe's rejections of causalism, psychologism, and the opacity of reasons cut deeply against the grain of current theorizing about acting for reasons. The dominance of causalism in particular seems only to have strengthened since her time, perhaps in part due to the growing influence of empirical psychology on the philosophy of practical reason. Some will doubtless see this as further evidence that we should dismiss Anscombe's views on acting for reasons as mysterious and unscientific. But this would be too quick. In reconstructing Anscombe's account of acting for reasons, I hope to have shown that she had deep and interesting reasons for holding it. Whether or not we ultimately accept a view like hers, it is worth treating it as a serious alternative to the views that have become current orthodoxy.

In particular, Anscombe's insights about acting for reasons far outstrip what Davidson took from them, and so we do not do her justice when we theorize about acting for reasons solely through a Davidsonian lens. Among other things, we risk underrating the work of one of the most important woman philosophers of all time in favor of the contributions of one of her male peers. Contemporary work in the philosophy of action takes Anscombe's work very seriously. If I have shown anything in exploring Anscombe's account of acting for reasons, I hope it is that the philosophy of practical reason should do the same.

[8] Many readers will probably find this part of Anscombe's view harder to swallow than her rejections of causalism and psychologism. While I cannot undertake a full exploration of its plausibility in this chapter, it is worth noting that its plausibility depends partly on whether recent work in psychology really has the upshot philosophers have thought it to have.

References

Alvarez, Maria (2010). *Kinds of Reasons. An Essay in the Philosophy of Action*. Oxford: Oxford University Press.

——— (2016). Reasons for Action, Acting for Reasons, and Rationality. *Synthese*:1–18.

Annas, Julia (1976). Davidson and Anscombe on 'The Same Action'. *Mind*, 85 (338):251–257.

Anscombe, G. E. M. (1957). *Intention*. Cambridge: Harvard University Press.

——— (1979). Under a Description. *Noûs*, 13 (2):219–233.

——— (2006a). The Causation of Action. In Mary Geach & Luke Gormally (eds.), *Human Life, Action and Ethics: Essays by G.E.M. Anscombe*. Exeter: Andrews UK Limited. pp. 89–108.

——— (2006b). Practical Inference. In Mary Geach & Luke Gormally (eds.), *Human Life, Action and Ethics: Essays by G.E.M. Anscombe*. Exeter: Andrews UK Limited. pp. 109–148.

Boghossian, Paul (2014). What Is Inference? *Philosophical Studies*, 169 (1):1–18.

Carruthers, Peter (2013). *The Opacity of Mind: An Integrative Theory of Self-Knowledge*. Oxford: Oxford University Press.

Cassam, Quassim (2014). *Self-Knowledge for Humans*. Oxford: Oxford University Press.

Dancy, Jonathan (2000). *Practical Reality*. Oxford: Oxford University Press.

Davidson, Donald (1963). Actions, Reasons, and Causes. Reprinted in: *Essays on Actions and Events*, 2nd ed., 3–19. Oxford: Clarendon Press (2001).

——— (1973). Freedom to Act. In Ted Honderich (ed.), *Essays on Freedom of Action*. New York: Routledge.

——— *(1984)*. Inquiries into Truth and Interpretation. *Oxford: Clarendon Press.*

Davis, Wayne A. (2005). Reasons and Psychological Causes. *Philosophical Studies*, 122 (1):51–101.

Ford, Anton (2017). The Representation of Action. *Royal Institute of Philosophy Supplement*, 80:217–233.

Gertler, Brie (2011). *Self-knowledge*. London: Routledge.

Ginet, Carl (2002). Reasons Explanations of Action: Causalist Versus Noncausalist Accounts. In Robert H. Kane (ed.), *The Oxford Handbook on Free Will*. Oxford: Oxford University Press. pp. 386–405.

Hursthouse, Rosalind (2000). Intention. *Royal Institute of Philosophy Supplement*, 46:83.

Hyman, John (2014). Desires, Dispositions and Deviant Causal Chains. *Philosophy*, 89 (1):83–112.

Korsgaard, Christine (2008). Acting for a Reason. In *The Constitution of Agency: Essays on Practical Reason and Moral Psychology*. Oxford: Oxford University Press.

Lord, Errol (2018). *The Importance of Being Rational*. Oxford: Oxford University Press.

Nichols, Shaun & Stich, Stephen P. (2003). Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds. *Oxford: Oxford University Press.*

Nisbett, Richard E. & Wilson, Timothy D. (1977). Telling More Than We Can Know: Verbal Reports on Mental Processes. *Psychological Review*, 84 (3):231–259.

O'Brien, Lilian (2015). Beyond Psychologism and Anti-Psychologism. *Ethical Theory and Moral Practice*, 18 (2):281–295.

Paul, Sarah K. (2011). Deviant Formal Causation. *Journal of Ethics and Social Philosophy*, 5 (3):1–24.

Sandis, Constantine (2015). Verbal Reports and 'Real' Reasons: Confabulation and Conflation. *Ethical Theory and Moral Practice*, 18 (2):267–280.

Schwenkler, John (2015). Understanding 'Practical Knowledge'. *Philosophers' Imprint*, 15.

Singh, Keshav (2019). Acting and Believing Under the Guise of Normative Reasons. *Philosophy and Phenomenological Research, 99 (2):409–430.*

Smith, Michael (1987). The Humean Theory of Motivation. *Mind*, 96 (381):36–61.

——— (1994). *The Moral Problem*. Oxford: Blackwell.

——— (2003). "Humeanism, Psychologism, and the Normative Story." *Philosophy and Phenomenological Research*, 67 (2):460–467.

Stoutland, Frederick (2011). Introduction: Anscombe's Intention in context. In Anton Ford, Jennifer Hornsby & Frederick Stoutland (eds.), *Essays on Anscombe's Intention*. Cambridge: Harvard University Press.

Thompson, Michael (2008). Naive Action Theory. In *Life and Action*. Cambridge: Harvard University Press.

Turri, John (2009). The Ontology of Epistemic Reasons. *Noûs*, 43 (3):490–512.

Vogler, Candace A. (2001). Anscombe on Practical Inference. In Elijah Millgram (ed.), *Varieties of Practical Reasoning*. Cambridge: MIT Press. pp. 437–464.

Wedgwood, Ralph (2006). The Normative Force of Reasoning. *Noûs*, 40 (4):660–686.

Wiland, Eric (2012). *Reasons*. New York: Continuum Press.

Wiseman, Rachael (2016). Routledge Philosophy Guidebook to Anscombe's Intention. *New York: Routledge.*

Part 3: The philosophy of practical reason as action theory and moral psychology

12. Three dogmas of agency theory

Nomy Arpaly

Philosophers who write about agency often subscribe to one or more of three views. The first view is that actions performed as a result of *deliberation* are somehow more "agential" than actions that are not the result of deliberation.[1] The second view is that actions *endorsed* by the person who performs them are somehow more "agential" than actions that are not endorsed by the person who performs them.[2] The third view is that actions with which the actor *identifies* are somehow more "agential" than actions with which the actor does not.[3] Actions that happen *without* deliberation, actions of which the agent *disapproves*, and actions from which the agent feels *alienated* are often looked upon as second-rate actions, deviant actions, or even non-actions – mere activities or behaviors. This is especially true for actions that fail all three tests, as in the following case provided by David Velleman:

> I have a long-anticipated meeting with an old friend for the purpose of resolving some minor difference; but ... as we talk, his offhand comments provoke me to raise my voice in progressively sharper replies, until we part in anger. Later reflection leads me to realize that accumulated grievances had crystallized in my mind, during the weeks before our meeting, into a resolution to sever our friendship over the matter at hand, and that this resolution is what gave the hurtful edge to my remarks. In short, I may conclude that desires of mine caused the decision, which in turn caused the corresponding behavior; and I may acknowledge that these mental states were thereby exerting their normal motivational force, unabetted by any strange perturbation or compulsion. But do I necessarily think that I made the decision or that I executed it? Surely, I can believe that the decision, though genuinely motivated by my desires, was thereby induced in me but not formed by me; and I can believe that it was genuinely executed in my behavior but executed, again, without my help. Indeed, viewing the decision as directly motivated by my desires, and my behavior as directly governed by the decision, is precisely what leads to the thought that as my words became more shrill, it was my resentment speaking, not I.
>
> (Velleman (2000:126–127))

Though my "accumulated grievances" in this example might be perfectly legitimate, and my resolution to break up with my friend a reasonable conclusion from them, I did not decide to break up with my friend *through deliberation*. It also seems to be the case that I *disapprove* of what I do while I do it: I do, after all, believe that my

[1] See, for example, Korsgaard (2009) and Raz (1999). Also see Arpaly and Schroeder (2014) for a critical discussion of the work of other authors.

[2] See, for example, Watson, as well as Korsgaard and Raz.

[3] The first defense of this view is in the work of Harry Frankfurt (1971, 1999), and Michael Bratman is another example (2007).

friend and I should simply resolve the "minor difference" we got together to resolve, and in addition, it is easy to assume I disapprove of my behavior because it is rude. Last but not least, I am *alienated* from my behavior: I experience myself not as doing something actively but as being passive with regard to something that happens to me – my voice turns louder, my words become shrill. I think of "my resentment" as taking over me and talking or, more sophisticatedly, of my desires causing a resolution which causes behavior – my desires and not I. Velleman suggests that when I feel that way, I am right: what happened was more than a mere movement of my tongue, but less than a full action.[4]

In previous work, I have argued against these three views. I have argued that the fact that a behavior is not the result of deliberation, is not endorsed by the person behaving, is experienced by that person as "external", or indeed all three at once, does not make the behavior in question anything less then an action. It need be no less of an action than anything an agent does as a result of deliberation, with full endorsement, and experiencing it as something *she* does (rather than her resentment or her desires). My rejection of the three views was very contrarian in its time[5] and is still quite controversial. In this article, I would like to sum up and restate this rejection and my arguments for it. Let us look at each view in turn.

[4] See also Raz (1999), chapter 1, and Korsgaard (2009), chapters 1 and 8.

[5] See Arpaly (2003).

Deliberation

To say that every action (or every action done for a reason) is the result of deliberation leads to an infinite regress, as deliberation itself is an action. If every action is the result of deliberation, deliberation itself is the result of deliberation, which is the result of deliberation, and so on ad infinitum. This, and other regress arguments against a variety of claims about the dependence of acting for reasons on deliberation, has been argued for in Arpaly and Schroeder (2012, 2014), but quite apart from these arguments, there is a simple fact to point out: if every human action, to be an action, had to be the result of deliberation, we would never get through the day.

Most of us need to deliberate when we plan a complex trip itinerary, calculate something using long division, or figure out whether to vote for a law that concerns matters we know little about. We do not, however, deliberate about all our actions. For example, if I want chocolate, I open the cupboard and take it from there. I do not, before opening the cupboard, need to perform the action known as deliberation – which in this case would amount to intentionally telling myself inwardly "OK, I want chocolate. What should I do to get some it?", focus my concentration on the subject, hunt for ideas, and come up with "ah! The cupboard!" This is the kind of thing we do when we try to figure out the solution to a hard problem, but not every time I *infer* something am I engaged in deliberation, and while there must be some kind of practical inference connecting my desire for chocolate, my belief about the contents of the cupboard, and my opening the cupboard, that inference was not obtained through an act of deliberating – not any more than I had to reflect to infer from believing that I just heard a loud "meow" to believing that my cat is being loud in the other room. Similarly, a basketball player does not have time to deliberate before he throws the ball to a teammate, yet throwing the ball is an action, and in the case of a talented player in a strategic game, it can be the result of a more complex – but equally spontaneous – practical inference.

It is relatively easy to see that the average tooth-brusher does not deliberate before brushing her teeth. However, it is important to remember that simple, habitual, or boring actions such as tooth-brushing are far from being the only actions that happen without deliberation. I have mentioned basketball, but the most striking example to me is conversation. A person like Oscar Wilde, who is in the habit of coming up with a witty response to any question or claim made by any interlocutor, does not, before responding, deliberate about the right way for him to respond. When Wilde bragged that he can make a pun on every subject, someone suggested the queen, whereupon Wilde responded very quickly – that is, with no time to deliberate – by

saying: "the queen is no subject". This was an action, and an action in response to some complicated reasons. Even conversations that do not involve this much wit usually consist of a sequence of speech acts that respond to reasons without the intervention of deliberation: deliberation has to come into play when we try, for example, to have a conversation with a person from a foreign culture whom we are trying not to offend, when we attempt to answer a question on a topic we find hard, or when we are not fluent in the language we use.

When acting without deliberation is mentioned by philosophers at all, it is often assumed that acting with deliberation is the default way to act and acting without it is the result of "shortcuts" that an experienced reasoner would develop over the years.[1] This might be a plausible theory of a very limited set of actions. Doctors who make decisions about medical procedures without a lot of deliberation were in fact once medical students who made the same decisions through deliberation and just "got better at it" over the years. However, in general we start out as acting but non-deliberating creatures (small children) and then become deliberating creatures, a fact that contrasts starkly with the idea of the agent who used to be a deliberator and then became more spontaneous through shortcuts. Unlike some actions learned in adulthood, many ordinary actions (opening the cupboard to retrieve chocolate being one) are actions that we never deliberated as to how to perform. Deliberation itself – remember? – is an action we have learned to perform, and we did not learn deliberation by *deliberating as to how to deliberate*, then developing shortcuts. Similarly, nobody learns how to talk in her native language through deliberation and reflection about how to talk – some amount of knowing how to talk seems to be there before one is old enough *for* deliberation, which might be, for many at least, a form of inwardly "talking to yourself". Similarly, the Oscar Wildes of the world do not learn to crack wise through deliberation, but rather their ability to crack wise starts early and develops with use. Usually, they cannot articulate anything like rules or best practices for the beginning wisecracker.

In truth, it is not spontaneous practical inference that is a "speeded up" version of deliberation, but rather deliberation is an action that we perform when the normal flow of spontaneous practical inferences is "stuck" or slowed down – perhaps there is an overwhelming amount of relevant data, perhaps it's too complex, perhaps we don't remember something right away, perhaps the task, (like long division for most of us) requires taking apart into smaller tasks that themselves are manageable without deliberation (like elementary arithmetic for most of us), perhaps there is a lot of distraction and we are trying to concentrate, and so on. Deliberation is a wonderful tool we use to solve such problems, but that is no reason to think that in the many occasions when the tool in question is not required – when the problems simply do not arise – then we are somehow not quite agents or not quite acting. This is evidenced in

[1] See Chan (1995:140 and fn 10).

the fact that we blame and morally credit people for undeliberated-upon acts all the time.

Endorsement

Human beings often do things that they think they shouldn't do – prudentially, morally, or all-things-considered. The literature on akrasia and weakness of will focuses disproportionally on cases involving dieting (occasionally exercise, drinking, and smoking), so it would be useful, I think, to run through our mind a sample of the immense variety of cases that exist of doing things that one thinks one shouldn't do. Consider a friend of mine who thought (rightly!) that he must not organize my severely ill-organized books without my permission but who did it anyway. Consider, more dramatically, a lesbian who has a sexual and romantic affair with another woman despite thinking that homosexuality is a grave sin and despite trying again and again to force herself to end the affair; a philosopher who repeatedly expresses his controversial opinions where he knows he should not; the adult who, against her best judgment and to her great frustration, acts in a childish manner whenever her parents are around; or the parent who cannot stop acting patronizingly around his adult son. In addition to these cases of stark akrasia or weakness of will, there are other cases of conflict between a person's views of what he should do and the way he routinely acts. Consider the person who goes to church, seems sincere as she talks to her daughter about the need to show charity and love one's neighbor, and then proceeds to work some extra hours at a proposal that the pharmaceutical company for which she works raise the price of a life-saving drug by 200 percent.

From Frankfurt (1971) and Watson ([1975]) on, it is common among theorists of agency to assume that the actions which we do not endorse – whatever that endorsement amounts too, exactly – are somehow less agential than others, or not actions at all, or not things that a person does freely.

Velleman seems to consider the case of the akratic yelling a case where it is not I, the agent, who is talking. A question that rises immediately is the question of my moral responsibility: am I not blameworthy for my rudeness to my friend? (Or, as my students like to ask, should my late paper be excused if, while playing computer games instead, I strongly believed that I should write it?) It seems as if I am normally blameworthy only for actions which I perform and not for any other type of event. Velleman replies in a footnote that I am blameworthy for my rudeness – indirectly. I have a responsibility to watch my emotions and not to allow, for example, my resentment to get out of hand. It is almost as if the emotions and desires of mine which I do not or would not endorse are like dogs or young children: I am responsible for their misdeeds in that I must control their behavior, but surely their misdeeds are not my actions.

How plausible it is to look at one's un-endorsed actions as analogous to the misdeeds of dogs and young children varies case by case, and some of the cases I mention are under-described. It seems most problematic, though, in cases where the agent seems wrong to disapprove of her action. Consider the case of the lesbian, whom I elsewhere called Lynn. (Arpaly 2003:16). Imagine that Lynn comes to her philosophy professor, even if he is an agency theorist, and tearfully confesses her romantic entanglement with the other woman. It is quite *un*likely (I hope!) that he will tell her that her sexual and romantic acts are not really her actions, that she should treat them as if they were the deeds of her runaway dog, or that it was only her romantic attraction acting, not she. The professor is likely to encourage her to accept *herself* the way *she* is.

But where the "runaway dog" view seems least plausible is in cases where the action which the agent performs despite his disapproval is a downright meritorious one. I have written in this context about Mark Twain's character, Huck Finn, who is a white boy in the antebellum south.[1] The uneducated Huck becomes a friend of a black slave, Jim. Keen on helping his friends, he wants to help Jim when Jim escapes slavery, but he assumes that it would be the wrong thing to do, struggles with himself, and resolves to turn Jim in. When opportunity comes, though, Huck cannot bring himself to turn Jim in and he shields him from the people who are looking for him. When Huck does so, it seems that he does so for morally good reasons, though he does not know that they are morally good reasons: the time he spends with Jim causes him to see Jim as a person just like him, and so the thought of turning him in becomes as painful to him as the thought of severely hurting any other friend would be. This can happen without Huck developing, even unconsciously, the counter-belief that it is morally good to help Jim – a belief that Huck has nowhere to get from and which he is not smart enough to come up with himself. Let us assume – in my view, consistent with the novel, but that should not matter – that Huck develops no such counter-belief.

If we assume that Huck cannot bring himself to turn Jim in as a response – to use a simplistic phrase – to Jim's personhood, his similarity to Huck, it seems that Huck is to some extent praiseworthy. He defies what he thinks of as morality rather than treating a human being and a friend in a certain brutal way. This is praiseworthy, just like my rudeness to my friend in Velleman's example is blameworthy. Velleman explains my blameworthiness by saying that what I am really blameworthy for is failing to control my emotions. What is Huck praiseworthy for, though? Surely not for controlling his emotions or for some character-building exercise he did when he was even younger. There is no course of action for Huck to be praiseworthy for except his shielding of Jim. Huck is not the only real or fictional person who practices *better* than he preaches and is judged mostly by what he practices and motivation for it, as an ethical person is not always a good theorist of ethics.

There are plenty of similar but less dramatic examples of praiseworthy un-endorsed action, Imagine a person who says "look, I don't really care about morality. I guess

[1] This case, first used to support a different view by Bennet (1974), was first used by me in Arpaly

I'm just not a moral guy: I drink, I smoke, I have sex without being married. What *do* I care about? I want people to live and let live, and not to be miserable. I realize I'd sometimes even do something immoral for these purposes". Imagine a case in which the right thing to do is to cheer up a miserable person through buying her a drink she likes and keeping her company. Our character takes drinking to be immoral, but if he were to buy the person a drink as part of ameliorating her misery, I take his action to be praiseworthy regardless of his thinking of it as vicious. There are also cases, not involving morality, in which we take an unendorsed action to be rational – or at least more rational than avoiding it would be (Arpaly 2000).

We hold people praiseworthy for actions that they do not endorse, and that suggests that we do not treat these actions as anything other than actions. We hold people blameworthy for actions that they do not endorse, and there seem to be no reason to speculate that there is somehow an asymmetry and some story about indirect responsibility is required to explain blame judgment that are analogous to the praise judgments. Actions that we don't endorse are ordinary actions, which dovetails with the fact that they are very common – the good, the bad, and the neutral.

and Schroeder (1999) and then in Arpaly (2003).

Identification

To identify with an action, one need not endorse the action, and, conversely, to be alienated from an action, one need not disapprove of it. As has been pointed out by Watson, it is possible to identify with features of yourself that you regard as vices. Imagine, for example, that Parvati believes that she spends more money on electronics than is prudent for her and often attempts to curb her spending. Still, when Parvati surrenders to temptation and buys the latest expensive gadget, shaking her head as she does so, she does not feel viscerally surprised, nor does she ask herself what got into her. Instead, she smiles sheepishly and says words that imply identification, words such as *here I go again* or *guess who absolutely had to spend $1000 on a new phone last night*. Or we can imagine Stan, who is addicted to alcohol and attempting to quit through Alcoholics Anonymous. Stan occasionally relapses, to his great shame and frustration, but instead of attributing his binge-drinking to possession by "the demon alcohol," he attributes it to his sinful self, in fact his *willfulness*. On the other hand, he attributes his increasing periods of sobriety not to his will but to *divine intervention*.[1]

So it is possible to identify with parts of ourselves that we do not endorse, such as imprudence or addiction.[2] It is also possible to feel alienated from parts of ourselves (or actions) that we do endorse. Stan does that. For another example, imagine that Svetlana normally feels confused, disorganized, and "bumbling", but then an emergency comes and she comports herself with efficiency and calm. Later, she says: "it was as if someone a lot more rational took over me and told me what to do. Then, when everything was ok, I reverted to my ditzy self". Svetlana, unlike Stan, does not believe in supernatural entities, but it sure *felt* like an alien presence when she was surprised by her own rationality.

So identification is not the same as endorsement and alienation is not the same as disapproval. It is also worth noting that occasions in which a person is alienated from his actions, desires etc. need not be occasions in which others regard him as "not himself", or vice versa. As I lose my temper akratically, feeling as if some demon took over my voice and gestures, my colleagues might be thinking "there she goes again. How typical!" It is perfectly possible for a person to be repeatedly alienated from his

[1] It does not matter here whether we regard addiction to alcohol as a disease. It is perfectly possible for a person to identify with it as if it were part of his "self" (as it is to be alienated from it). If one can identify a disease as part of a self, it is just another testimony to the thesis I'll defend later: that what we identify with depends on our visceral self-image and not on any important truth about who we are.

[2] This is pointed out by Watson in a later paper (1977).

fits of anger while his friends and colleagues think of these fits as him showing his "true colors".

If identification does not correlate with endorsement and alienation does not correlate with disapproval, what *do* identification and alienation correlate with? As Schroeder and I argued (Schroeder and Arpaly 1999), we feel alienated from behaviors and motives when they are opposed, more or less dramatically, to our visceral self-image, and we identify with behavior that fits that image. Compare the way a person who still viscerally sees herself as young can viscerally feel as if the new patches of white hair she sees in the mirror were painted onto her as some kind of joke, or a person who still viscerally sees himself as poor can feel like an impostor as he enters his expensive car. Something similar happens when a person who sees herself as sweet and kind feels a surge of anger that conflicts with that self-image: the anger can feel to her like something external "possessing" her. If she saw herself as an aggressive person from the start, the anger would have felt like part of her, even if not her favorite part. Thus, if Parvati felt that excessive love of gadgets is an affliction that only men suffer from, or if Stan felt that alcoholism is a working-class phenomenon while identifying with his position as a high-end lawyer, they *would* have felt alienated from their habits, after all. Though it might surprise those of us who experience alienation often, some people experience it rarely or not at all, not because they endorse all their actions and desires but rather because their self-images are more realistic, or vague, or open in the spirit of the saying "nothing human is foreign to me".

What, then, is the significance of identification and alienation? Though they are fascinating phenomena, there is little reason to believe that they are highly significant to the specific matter of agency. The person alienated from her relatively new patches of white hair has to admit, if she is not psychotic, that the white hair belongs to her, and no lack of identification will make it black again. The person alienated from his expensive car can reassure himself through looking at his documents that the car belongs to him, and his sense of being an impostor has no bearings on its legal status. Similarly, feeling that your angry action is not really yours does not make it less yours. Given the general unreliability of visceral self-images, it would be very strange if it turned out that the status of each of my actions as "really my action" or less so lines up squarely with whether or not *I experience it* as such – especially if the way I experience it is influenced by such silly notions as "only men buy electronics" or "only working-class people get addicted" or, for that matter, "I never get angry".

Conclusions

Sometimes, we do things we don't deliberate about, things that we do not endorse, or things that we do not identify with. Even in Velleman's case, in which what I do is neither the result of deliberation on my part nor something I endorse nor something I identify with, it is I who yell at my friend and my yelling is an action. To see that most clearly, one need only perform the Huck Finn exercise and imagine that my yelling was a meritorious thing to do. Perhaps the person whom I took to be my friend has in fact been exploitative and demeaning to me, perhaps due to his sexism. Without deliberation, I have gradually noticed the demeaning side of his behavior and my resentment has built up over time. Still consciously underestimating the badness of his behavior, I attempt to discuss "a minor difference" with him, but deep inside I have finally had it, and so I hear myself telling him loudly to go to hell. As I leave, I feel liberated, if shaken, and my friends, who have worried about the nature of my relationship with the person in question for a while, are happy that I finally "came to my senses". It would have been even more effective for me, they are willing to admit, if I left the conversation and the relationship with the man I took to be a friend in a more dignified manner, making a devastating remark instead of a shrill one, and perhaps, I not being a "natural" for this kind of thing, this could have been accomplished only if I had deliberated correctly about the state of the relationship and planned my "exit" in advance. However, admitting that I could have done it more gracefully does not amount to saying that my shaking off a person who treated me despicably was not my action. It is something that I deserve to feel good about – depending on the background, even *proud of* – in a way that one deserves to feel only about *one's own action.*

Why, then, the impression that so many philosophers have that actions that are not the result of deliberation, that we do not endorse and with which we don't identify are not actions or are "less" actions?

One answer that comes to mind would invoke an expression used by Velleman: human agency *par excellence.* When we deliberate, endorse, or identify, we exhibit abilities that only humans have and might even seem an essential part of the state of being an adult human being. However, the fact that the abilities in question are, as far as we know, uniquely human does not in itself imply that we are "more agential" or more rational when our actions are related to these abilities. Some people have the ability to play basketball well, and many people cannot do so. That does not mean that, for those who can play basketball, their actions on the basketball court are somehow more "agential", more "action-y", or more rational than their actions outside

the basketball court. They are interesting and complicated actions that others cannot perform – and that's all.

In a related way, some of us – interestingly, not all of us – would find it sad if we found out that the only way in which humans are superior to other animals is in being more cognitively sophisticated. Such people would prefer to think of humans as possessing a quality – rationality or "autonomy" or "rational autonomy" come to mind – which sets us radically apart from other animals, and identifying agency with specifically human capacities can be tempting for this reason. If I am an agent and my cat is not, the divide between us is indeed radical. However, whether it is sad or not, I think it's quite misleading to speak of a radical divide like this. This might seem like a strange thing to say given the old Aristotelian thought that we, and we alone, are rational animals, but I think this thought is somewhat misguided. I am doubtless more *cognitively sophisticated*, by far, than my spotted cat, Gaius Valerius Catullus, but arguably he is at least as *rational* as I am, if not more so, good as he is at obtaining the means to his ends and impressive as he is in never attempting a high jump unless he can complete it safely. If you do not immediately grasp the idea of A being less cognitively sophisticated than B but A being more rational than B, consider for a moment Dolores, a mentally healthy girl of 11, and then consider again the same girl at age 13, having passed puberty and acquired the social role of a teen. If Dolores' parents are the sort of people who use the word "rational", they would be likely to lament how irrational their daughter has become in comparison to her two-year-ago self. Yet, it is simply true that, assuming she had grown up normally, Dolores the somewhat troubled adolescent is more intelligent – and thus more cognitively sophisticated – than Dolores the perfectly reasonable, well-adjusted child. Any advantage I have over my spotted cat is not due to a binary difference in rationality – Nomy is rational, Catullus is not – but to a large difference-in-degree with regard to cognitive sophistication – Nomy can process and understand more numerous and complicated data than Catullus can.

Another reason the three dogmas might be tempting has to do not with our relation to other animals but with the dominance, in the life of most academics, and many others, of inner struggles involving difficulties sticking to resolutions, also known as problems with self-control. Two examples immediately come to mind. First, many "first world" people would like to eat less than they do, either for health reasons or because of puritan norms and gender roles, and the related inner struggle can occupy one's mind quite a bit. Second, many writers, including academics, face a daily fight with procrastination and are always convinced, rightly or wrongly, that they are working too little. A writer's inner struggles between the need to write and temptations to do other things can also weigh heavily on her mind. In general, of the very many causes of irrationality (think of people whose behavior is twisted by anxiety, depression, wishful thinking, paranoid ideas, bigotry, plain old fatigue, or the mad resolve of anorexia) failures of *will power* are perhaps the most salient from the introspective point of view because one *knows* or believes that one is irrational at the time of the failure (other irrationalities are mostly noticeable in others). When such failures are salient, it is

tempting to see some kind of self-government (or "autonomy") as the be-all and end-all of rationality or agency. This in turn can result in thinking of yourself as country that needs to be well governed, your deliberation as the deliberations of the inner congress or parliament, un-endorsed and surprising actions as severe disturbances of the peace, and some desires as "outlaw desires". It is a natural way to imagine things, to be sure, but natural metaphors can be very misleading. Think, for example, of the idea that an angry person needs to "let off steam" (in a safe place, of course) or she might explode. Therapists believed in the dangers of "pent up anger" for years, but the scientists have thoroughly debunked the picture of people as kettles. What is normally called "letting off" anger reliably *increases* it. The "self-government" metaphor for rationality and agency need not, at the end of the day, be any more accurate, and it is my suspicion that people are not any more like countries than like kettles. The "government" metaphor has inspired great work, but I suspect that ultimately is not, as it were, where the action is.

References

Arpaly, N. 2000. "On Acting Rationally Against One's Best Judgment." *Ethics* 110: 488–513.

Arpaly, N. 2003. *Unprincipled Virtue*. Oxford: Oxford University Press.

Arpaly, N. and Schroeder, T. 1999. "Praise, Blame, and the Whole Self." *Philosophical Studies* 93: 45–57.

Arpaly, N. and Schroeder, T. 2012. "Deliberation and Acting for Reasons." *Philosophical Review* 121: 209–239.

Arpaly, N. and Schroeder, T. 2014. *In Praise of Desire*. Oxford: Oxford University Press.

Bennett, J. 1974. "The Conscience of Huckleberry Finn." *Philosophy* 49 (188): 123–134.

Bratman, M. 2007. *Structures of Agency*. Oxford: Oxford University Press.

Chan, D. 1995. "Non-Intentional Actions." *American Philosophical Quarterly* 32: 139–151.

Frankfurt, H. 1971. "Freedom of the Will and the Concept of a Person." *Journal of Philosophy* 68: 5–20.

Frankfurt, H. 1999. "Identification and Externality." *Philosophical Review* 86 (3): 316–339.

Korsgaard, C. 2009. *Self-Constitution*. Oxford: Oxford University Press.

Raz, J. 1999. *Engaging Reason*. Oxford: Oxford University Press.

Schroeder, T. and Arpaly, N. 1999. "Alienation and Externality." *Canadian Journal of Philosophy* 29 (3): 371–388.

Velleman, J. D. 2000. *The Possibility of Practical Reason*. Oxford: Oxford University Press.

Watson, G. 1975. "Free Agency." *Journal of Philosophy* 72: 205–220.

Watson, G. 1977. "Skepticism about Practical Reason." *Philosophical Review* 86 (3): 316–339.

13. Some reflections on the relationship between reason and the will

Sarah Buss

"Practical" reasons are what theoretical reasons are not. It is surprisingly difficult to offer a positive characterization. According to some, a reason is practical insofar as its significance can be captured by reasoning that concludes in intentions. But there appear to be reasons for emotions and desires, and for judgments regarding what it makes sense to do. Are not these reasons practical, too?

Whatever the answer to this question, it is widely agreed that in many cases, there need not be a very tight connection between (i) an agent's responses to her circumstances and (ii) her apprehension of reasons. In particular, a person can feel or desire something, even if she thinks that her reasons for feeling or desiring otherwise are much more compelling. But is this sort of mismatch possible for every sort of response? In particular, is it possible for someone's intentional actions to conflict with her normative assessments of these actions?

Most philosophers take it to be obvious that a person can intentionally, and even "freely," *do* something even while she believes – and is well aware she believes – that she has more compelling reason to do otherwise. This, they claim, is the possibility of a paradigm form of irrationality (hereafter, "the Alleged Possibility"). I am among a small minority who reject this possibility. In my view, irrationality in this form cannot be reconciled with a fact so basic as to be a truism: an instance of behavior cannot be a person's action if it is the effect of psychic forces whose motivating force she herself opposes.

Having briefly explained what this truism implies about the relation between reason and the will, I will draw a more general lesson concerning the metaphysical constraints on irrationality. I will then turn my attention to the fact that rational beings appear to be able to exercise their agency in ways that do not depend on employing their reason. A familiar example involves someone who does not apprehend any reason to favor one of two alternative courses of action over the other. Under these circumstances, the person takes herself to have compelling reason to start down *one path or another* but no reason to start down *this path rather than the other*.

What does a choice amount to in this sort of case? How does *intentionally* setting off down *this particular path* differ from discovering that one is beginning to do so? I do not have a satisfactory answer to this question. I hope, however, to shed some light on the relation between reason and the will by exploring several cases in which someone's agency is dissociated from her capacity to reason. Though the agents in these cases are not pushed or pulled "against their will," there is an important sense in which they are passive in relation to their own behavior: in their capacity as beings who form opinions about what makes sense, they are bystanders to the causal power of their impulses; and so, they are bystanders to their own agency.[1] Such passive agency is

[1] In his influential papers on free will and moral responsibility, Harry Frankfurt fails to distinguish between being a passive victim and being a passive bystander. The person who fails to act "of his own free will" is, Frankfurt says, "conquered" by his strongest desire (Harry Frankfurt, "Freedom of the Will and the Concept of a Person," in *The Importance of What We Care About* (New York: Cambridge University

possible because even though no one can wittingly defy the verdicts of her own reason, there is more to the capacity to will than the capacity to put reason to practical use.

Press, 1988, pp. 11–25)), "hopelessly violated" by it ("Freedom of the Will," p. 17), "stampeded" ("Three Concepts of Free Action," in *The Importance of What We Care About* (New York: Cambridge University Press, 1988, pp. 47–57)). But Frankfurt also refers to this agent as a "helpless bystander to the forces that move" him ("Freedom of the Will," p. 21 and "Three Concepts of Free Action," p. 54). As I hope will be clear by the end of this chapter, this mixing of metaphors obscures important metaphysical distinctions.

The conditions of action as limits on the possibility of practical irrationality

A cow cannot wonder whether she has sufficient reason to move her head downward in order to bite off a bit of grass. A toad cannot wonder whether she has sufficient reason to hop down the road. Indeed, nonrational animals cannot wonder whether they (really) have *any* reason to do what they do – even if there is some sense in which their environments appear to them to give them reason to do one thing rather than another. This condition has some obvious disadvantages. But among the benefits is mental harmony: because a nonrational being cannot reflect on the desirability of its own impulses, it cannot occupy a point of view that is distinct from these impulses. As has often been noted, the capacity for self-evaluation grounds a vulnerability to self-alienation.

To acknowledge this vulnerability is not yet to determine its extent. In particular, we cannot assume that a rational being can oppose her own action at the very time she is engaged in acting this way. So we cannot assume that a rational being qualifies as an agent as long as she satisfies the conditions under which a cow and a toad qualify as agents.

Whereas it goes without saying that cows and toads are not opposed to being moved to do what their impulses move them to do, this does not go without saying for creatures like us. Because we can reflect on the desirability of being moved by our impulses, whether our behavior qualifies as our action depends on more than whether it can be directly attributed to our psychological states. Nor does it suffice that this behavior reflects our assumptions about how to achieve whatever ends are associated with these states. An adequate account of our kind of agency will distinguish cases in which *we* initiate our behavior from cases in which, however instrumentally rational we may be, the ends served by our behavior are not our own. This means that an adequate account of rational agency will distinguish (i) cases in which someone's desires move her to act in certain ways from (ii) cases in which her desires could move her to do something only if they could prevent her from acting.

A rational being is opposed to doing something when she believes that, all things considered, she has overriding reason not to do it. The point is not that someone necessarily opposes whatever action would be worse according to the standards of

evaluation she accepts. (For the purposes of this discussion, I am happy to allow that someone can believe she has sufficient reason to do something, even though she also believes that acting otherwise would be even better.) The point is that, if someone concludes that she has more compelling reason not to do something, then she is opposed to doing it. More carefully, this evaluative judgment constitutes *her opposition* if and only if it reflects a practical aim – the aim of determining how she will act. Lacking such a practical aim, an agent would not be the least bit conflicted, even if she did not endorse her behavior. Even if she thought that the reasons against behaving this way "outweighed" the reasons in favor of behaving this way, if she did not care how she acted, then her behavior would not be contrary to what she was committed to doing.[1]

I will take a closer look at this sort of self-detachment in the second half of the chapter. For now, I want to stress that it is not the sort of self-relation that characterizes someone who wittingly occupies a point of view that is opposed to her own intentional action. Like anyone who does things for reasons, such a person (the "clear-eyed akratic agent" *as she is typically characterized*) has considered the pros and cons of various possible actions *in order to determine which option she will pursue.*[2] She has weighed her reasons because *she is committed to being guided by her normative/evaluative verdicts* (as a means to the end of doing what she has reason to do). Of course, at some point before she begins acting, she could lose interest in being responsive to

[1] It may seem that a regress threatens: I stress that the aims of *our desires* need not be *our own* aims, yet I identify the aim of determining the will as the key to whether a normative judgment is incompatible with willing otherwise (as opposed to merely being the judgment that we have compelling reason not to will otherwise). Could someone who draws her own conclusions about what she has reason to do be forced ("against her will") to be moved *by these conclusions*? This would be possible only if in order for the direct effect of an agent's normative judgment to be attributable to her, she would have to take herself to have sufficient *reason* to be moved by this judgment. But, clearly, we could ask the same question about this additional normative judgment. As I note in the chapter's opening pages, I do not have a satisfactory account of the agency of rational beings when reasons run out. What I can say, however, is that if someone thinks it would be better *not* to behave in a way that is responsive to her reason (at least on this particular occasion), then whatever she does will, necessarily, reflect *this* normative judgment. If, on the other hand, she forms no opinion one way or the other (as is surely usually the case in well-functioning rational agents), then this is another respect in which, as I note at the end of this chapter, we are always passive in relation to our own agency: when we are functioning well, we simply find that we are moved by our normative judgments; this is not up to us. In effect, what is true of a nonrational animal's relation to all her desires is true of a rational animal's relation to her disposition to conform her actions to the verdict of her reason.

[2] There are many interesting reflections on the relationship between *determining what to do* and *determining what is worth doing.* Gary Watson nicely summarizes the assumption that we assess our options *in order to determine our will*: "Practical reasoning is reasoning about what is best (or satisfactory) to do *with a view to making up one's mind about what to do.*" It is reasoning with the aim of "making a commitment to a course of action *by making a judgment about what is best (or good enough)* to do"("The Work of the Will," in *Agency and Answerability*, Oxford: Clarendon Press, 2004, pp. 123–157). In a recent paper, Matthew Silverstein defends a different – more externalist – picture of the relation between reasoning about what to do and reasoning about what we *ought* to do. See Matthew Silverstein, "Ethics and Practical Reasoning," *Ethics* 127, no. 2 (January 2017): 353–382.

the reasons she discovers. But if this does not happen – if she maintains her aim of employing her reason in order to determine her will – then anything she might do which was, by her own lights, incompatible with this commitment would, by her own lights, be something she was opposed to doing. Under these conditions, she would almost surely think that her behavior had *something* going for it. Nonetheless, by stipulation, she would not think there was *good enough* reason for her to behave this way. So, whatever forces moved her to behave this way would thereby prevent her from initiating the behavior herself. In short, they would prevent her from acting.

This train of thought appeals to the fact that the aim of employing one's reason in order to determine one's will is not just any old aim. It is, rather, an aim concerning the conditions of one's own agency. This is why in doing something one takes to be contrary to this aim, one would not just be doing something that conflicts with one or more of one's desires. One would be failing to comply with the conditions one has imposed on one's own agency. To aim at being constrained in one's actions by one's own normative verdicts just is to regard the pursuit of any other aim as incompatible with one's agency. If one's rationale for reaching a normative verdict is to settle the question as to which of one's desires to satisfy, then, given that none of these desires is itself a response to this question, none of these desires could move one to defy one's own normative verdict without pushing or pulling one to do something "against one's will." To put the same point the other way around, whereas one can continue to desire to do something even if this desire is not expressed in one's action, if, by one's own lights, one's judgment about what it makes sense to do right now owes its existence to its capacity to determine one's action, then one's action cannot fail to reflect this judgment unless one has, as we say, "changed one's mind."

We can approach the same point from a slightly different angle by considering the nature of *practical* reasoning. A problem with regarding normative judgments as the conclusions (or at least the initial conclusions) of practical reasoning is that, as I have suggested, we can form such judgments without having any commitment to acting. But the *aim* of acting does not seem to be the essential feature either. After all, we can engage in *theoretical* reasoning for a practical purpose: I can do a math calculation in order to determine how much sugar to put in the cake, in order to decide which route to take to my destination, or just in order to impress a friend. This has led some philosophers to suggest that the aim of practical reasoning must somehow be "internal to its own content."[3] The point, I take it, is that the fact that one is reasoning *in order to figure out what to do* must somehow figure in the significance one attributes to each premise. Without wading too far into these deep waters, I want to suggest, in particular, that if *instrumental* reasoning is not merely *causal* reasoning, this must be because it involves identifying causal relations under the guise of the aim of determining what

[3] Mueller, Anselm, "How Theoretical is Practical Reason," in *Intention and Intentionality: Essays in Honor of G.E. M. Anscombe*, ed by Cora Diamond and Jenny Teichman (Sussex, England: Harverster Press, 1979, pp. 91–108).

is worth doing. An episode of reasoning qualifies as an episode of *practical* reasoning (rather than simply reasoning from effect to cause and cause to effect) insofar, and only insofar, as it is guided throughout by the reasoner's assumption that she is determining what-is-worth-doing-in-order-to-do-it. According to this suggestion, for a practical aim to be "internal" to the significance of each premise just is for each premise to owe its significance to the reasoner's aim of determining-what-to-do-by-determining-what-is-worth-doing.[4] But this means that if someone's reasoning about how to achieve some end is not merely theoretical reasoning about what causes what – if it is reasoning she undertakes in order to determine her will – then it is essential to this reasoning that the conclusion it generates regarding what, all things considered, she has reason to do reflects what she is committed to doing. This means that, as long as she does not revise this conclusion, she is opposed to behaving otherwise. So, were she to behave otherwise, this behavior would not qualify as her action.

This critique of the Alleged Possibility points to limits on the possibility of unfree, as well as irrational, action. In particular, it implies that most popular philosophical accounts of "compelled" or "compulsive" agency must be rejected along with the Alleged Possibility. If it is not possible *freely* to intentionally and wittingly defy one's normative judgments regarding one's present course of action, this is because it is not possible to act in (avowed) defiance of one's normative judgments regarding one's present course of action, whether freely or unfreely.[5] A person can, of course, be moved to do something by a desire she wishes were far weaker than it is. The point is simply that there are metaphysical limits on how this can happen. If a person's desires can interfere with the role that her reason plays in determining her actions, this must be because they can interfere with her reasoning. It cannot be because they can cause her intentionally to do things even while she is aware that her own normative verdicts do not support doing them – at least not if these verdicts reflect her interest in here and now determining her will.[6] The point is really very simple: because no desire or combination of desires

[4] I think Mueller fails to capture the necessary content when he refers to "my considerations being conducted with a view to [my] end" (p. 99) What is internal to instrumental reasoning, insofar as it is practical, is not just that I am considering possible actions with a view to some end, but that I am considering them with a view to their *instrumental relation* to this end – that is, not merely as possible causes, but as *means*, that is, as something it *makes sense to do in order to achieve* my end.

[5] For an important discussion of weakness of will that explicitly acknowledges the impossibility of distinguishing weak-willed agents from compulsive agents when the former are characterized as wittingly doing what they take themselves to have decisive reason *not* to do, all things considered, see Gary Watson, "Skepticism about Weakness of Will," *The Philosophical Review* 86, no. 3 (July 1977): 316–339. My point is that in addition to obscuring the distinction between weakness of will (a failure of rationality) and compulsion (a failure of self-determination), such accounts of weakness of will cannot do justice to the distinction between acting unfreely and not acting at all.

[6] For a detailed exploration of both the extent to which even paradigm nonautonomous agents – for example, those diagnosed with OCD or anorexia – endorse the working of their will, and an alternative account of their lack of autonomy, see Sarah Buss, "Autonomous Action: Self-Determination in the Passive Mode," *Ethics* 122, no. 4 (July 2012): 647–691.

is a stand-in for the agent herself, more is required in order for someone to act than for her to be moved to do something by one or more desire to do it.

When an agent takes herself to be opposed to doing the very thing that something is now forcing her to do, the effects of this force cannot be attributed to the agent herself. This is no less true when the force operates on her by operating *in* her. If some internal mechanism were somehow to cause a rational being to hop, even though she was convinced that hopping was the last thing in the world she had reason to do, and if she formed this conviction in answer to the question whether to hop, then when she hopped, it would be as if – against her will – certain powerful electro-magnetic forces had gained control over the movements of her legs. The comparison would be no less apt if she were convinced that her legs would not have moved in this way if there had not seemed to her to be at least something appealing about hopping in this way. Under these circumstances, too, *her hoppings* would no more qualify as *her actions* than the movement of her leg qualifies as her action when it occurs in direct response to the hammer's tap on her knee. Someone cannot turn her reflexive behavior into an action simply by noting that there is something to be said for behaving this way. Nor will it suffice for her to be disposed to adjust what she does whenever this adjustment is necessary in order to achieve the end at which the behavior is directed.

Some philosophers will insist that in contrasting rational agents with cows and toads, I have failed to acknowledge an obvious possibility. If a person can act for reasons, these philosophers claim, then she has the capacity to *choose* or *decide* to act in a way that conforms to the verdicts of her reasoning. But, they add, the capacity to choose or decide to do something is inseparable from the capacity to choose or decide *not* to do it. Accordingly, the capacity to choose or decide to do something is – necessarily – the capacity to exercise one's agency in a way that is not constrained by one's reason.

According to this suggestion, a rational agent can wittingly oppose herself to her own reason. I have already explained why this is not conceptually possible. But we can press the point by asking: what, exactly, could such defiance amount to, if the verdict of the agent's reason reflects an aim that she herself does not disavow – the aim of being constrained by the verdict of her reason? If an agent has not abandoned this aim, then if she believes that she lacks sufficient reason to do X now, she continues to be committed to not being moved by whatever desires she may have to do X now. Again, it is this aim that is expressed in her normative verdict. But then she cannot defy this verdict without, at one and the same time, committing herself to doing X now and being committed to *not* doing X now. This is how she herself would have to interpret her situation. But a person cannot wittingly believe that she is committed to doing this-very-thing (X)-that-I-am-committed-to-not-doing. A commitment to doing this very thing depends on a belief in the possibility of doing it. So, if someone believes that doing this very thing, X, is incompatible with doing something else, Y, then she cannot wittingly sustain a commitment to doing X now, even while being committed to doing Y now.

There is at least one more way to challenge the conception of the power to will as the power to self-consciously defy one's own reason: we can demand an account of what distinguishes someone who allegedly exercises this power from someone who is passive in relation to its manifestation. To say that an exercise of will is "intrinsically agential"[7] is not to answer this question; it is not to show that the looked-for distinction can be found. Until an account is forthcoming, we are thus forced to conclude that we can no more oppose our own all-things-considered normative judgments when, by our own lights, they are our means of determining what we will do than we can oppose our own will. The capacity to choose is the capacity to put our reason to practical use, and the capacity to put our reason to practical use is the capacity to impose volitional limits on ourselves. We can remove these limits in the same way that we impose them (by reaching a new normative verdict). We can also – if only for brief periods of time – dissociate these verdicts from any practical aim. There is, however, nothing we can possibly do to defy them.

[7] I am here referring to the conception of the will defended by R. Jay Wallace. As he puts the point, decisions and choices are "primitive examples of the phenomenon of agency itself." (R. Jay Wallace, "Three Conceptions of Rational Agency," in *Normativity and the Will*, Oxford: Oxford University Press, 2006, pp. 43–62) Wallace defends his view by appealing to the possibility of akrasia. "There has to be something in the act of choice," he says, "that distinctively goes beyond normative commitment if we are to leave room for akrasia and other forms of irrationality to which action is characteristically subject." (R. Jay Wallace, "Normativity, Commitment, and Instrumental Reason," *Philosophers' Imprint* 1, no. 3 [December 2001]: 1–26.) This point is similar to the point Kant makes in arguing that rational agents must be capable of the sort of radical choice that determines whether they treat the Categorical Imperative as a constraint on their choices. (See Kant's discussion of the distinction between *Wille* and *Wilkuer* in Immanuel Kant, *Religion Within the Limits of Reason Alone,* trans by Theodore Greene and Hoyt Hudson, New York: Harper & Row, Publishers, 1960, Book I, especially p. 20 and p. 38.) The position I am here defending is, in effect, a response to Wallace's challenge. In addition to explaining why I reject the Alleged Possibility, which Wallace and so many others accept, I hope to show that *intentions* can "go beyond" normative commitments without "going against" them and that it is thus possible to act contrary to one's own best judgment without manifesting a "weak will" of the sort at issue in discussions of clear-eyed "akrasia."

Further thoughts about the metaphysical limits on irrationality

In rejecting the Alleged Possibility, I have argued that the necessary conditions of agency impose limits on the ways it is possible for someone to be irrational. Before I turn my attention to the fact that rational agents can do things without being guided by their reason, I want briefly to note that the sort of constraint to which I have called attention is an instance of a more general metaphysical constraint on the ways in which someone can be at odds with herself.[1] In a nutshell: the limits on the possibility of irrationality are limits on the possibility of disunity. A single person would have to break in two in order to walk in two different directions at once. And no single person is capable of occupying a position that is above or below where she is. So, too, it is not possible for someone to occupy a *mental* point of view that is avowedly at odds with itself.

To appreciate this fact is to appreciate that the allegedly paradigm principles of rationality do not spell out *normative* constraints on the formation of beliefs and intentions.[2] Rather, they indicate *metaphysical* constraints on which combinations of attitudes a single person can wittingly attribute to herself. I want to illustrate this point by shifting my attention for a moment to a second principle that is relevant to the relationship between reason and the will – as relevant as the alleged requirement to avoid doing what we believe we lack sufficient reason to do.

To see that this second principle does not spell out normative constraints on willing is to be reminded of what we have already learned from examining the first principle: our capacity to do things for reasons constrains our capacity for irrationality because, and only because, it is a capacity to set ends and because there are (nonnormative) limits on which combination of ends we can have.

---

[1] For a lengthy discussion of the point to follow, see Sarah Buss, "Norms of Irrationality and the Superficial Unity of the Mind," unpublished manuscript.

[2] This is why they do not tell us what we have reason to do. For a sample of the many articles on whether we have reason to be rational see John Broome, "Does Rationality Give us Reasons," *Philosophical Issues* 15 (2005): 321–337), John Broome, "Is Rationality Normative?" *Disputatio* 2, no. 3 (November 2007): 161–178), and Niko Kolodny, "Why Be Rational?" *Mind* 114, no. 455 (2005): 509–563. For a discussion of the same issue in a very different tradition, see Christine Korsgaard, "The Normativity of Instrumental Reason," in *The Constitution of Agency* (Oxford: Oxford University Press, 2008, pp. 27–58). For some recent books on the subject, see Benjamin Kiesewetter, *The Normativity of Rationality* (Oxford: Oxford University Press, 2017); Errol Lord, *The Importance of Being Rational* (Oxford: Oxford University Press, 2018); Alex Worsnip, *Fitting Things Together*, unpublished manuscript.

To see that this is a very general point, let us turn, then, from *defying one's own reason* to *wittingly refraining from willing the means to one's end*. There is almost universal agreement that we are rationally required to will to take the means necessary to achieving our ends. Indeed, the principle of "instrumental rationality" is assumed to be the paradigm case of a rational requirement. If, however, as is surely the case, to commit oneself to *achieving* a given end *just is* to commit oneself to *doing what one must do in order to achieve* it (under this description), then it is simply *not possible* wittingly to do the very thing that the principle claims it would be *irrational* to do. Of course, it is all too possible to march off in a southerly direction under the mistaken impression that one is heading north. More generally, it is possible to be mistaken about what sort of things one must do in order to achieve one's ends. But ignorance is not the same thing as irrationality. And, importantly, the principle of instrumental rationality does not amount to the useless advice to avoid being ignorant or the equally useless advice to avoid making a mistake.

Of course, a person's self-interpretations can be more or less plausible. They can, in other words, be more or less responsive to the evidence – more or less responsive to reasons. But, again, being more or less rational *in this sense* is not what is at stake in the principle of instrumental rationality – nor does anyone suggest otherwise. The principle of instrumental rationality no more tells us to be responsive to reasons than it tells us to avoid being ignorant. It does not tell us to avoid telling convoluted – reasons-unresponsive – stories about why our actions really are perfectly compatible with our ends.

What, then, *does* the principle of instrumental rationality tell us? It simply calls our attention to the relation between willing the end and willing the means. In so doing, it reminds us of a constraint on self-interpretation: if you are a single person, then you cannot regard yourself as having an end without regarding yourself as being committed to taking the necessary means (unless, of course, you do not know what it is to have an end). Given that no single person can refrain from committing herself to taking the (acknowledged) necessary means (so conceived) to her avowed ends, if someone is, in fact, a single person, then she cannot refrain from attributing to herself the commitment to taking the means to her ends. The question is simply whether her self-interpretation satisfies this conceptual constraint in a way that is responsive to reasons.[3]

Again, this is a very general point: there are constraints on the ways someone can be irrational because there are constraints on which (combinations of) commitments it is possible for someone to attribute to herself. Notice that these limits do not prevent us from doing what could be properly described as acting "contrary to our best

[3] The most important point is this: *either* willing the end involves willing the means, in which case the principle of instrumental rationality does not spell out a normative constraint, *or* I am mistaken about the relationship between willing the end and willing the means, in which case the rational requirement indicates a *substantive* norm. In "Norms of Rationality," I note that the same basic point applies to formal principles of rationality governing the relations of beliefs.

judgment."[4] After all, we can always find reasons to reject a normative judgment that is at odds with a strong desire (e.g., "It is just this once," "I don't feel like being a slave to my reason right now," "I've been so good; I deserve a small reward," etc.).[5]

When we rationalize in this way, we put a necessary condition for the possibility of reasoning to a reason-undermining use. In order to reason about what to do, we must treat some facts as reasons for doing some things and not others *without reviewing the reasons for and against treating these facts in this way.* This means that it is possible for us to treat some facts (e.g., "It's just this once") as reasons for doing some things and not others, even while suspecting that we would not like what we would discover were we to review the reasons for and against treating the facts in this way. (This is the key to the double consciousness we experience when we experience ourselves as weak willed.)

*

[4] Nor do they prevent us from knowing that we are incoherent in certain ways. This, I take it, is the lesson of the Preface Paradox. More importantly – and sticking with reason in its *practical* capacity – almost all of us have a range of very broad commitments (e.g., to being good enough parents, friends, teachers, citizens, etc.) that we know could ground incompatible requirements in certain unforeseen circumstances. Not only, moreover, is such incoherence perfectly possible; there is nothing irrational about being in this condition. (For more on this point, see Sarah Buss, "Moral Requirements and Permissions, and the Requirements and Permissions of Reason," in *The Many Moral Rationalisms*, ed. by Karen Jones and Francois Schroeder, Oxford: Oxford University Press, 2018, pp. 110–144, "Personal Ideals," unpublished manuscript.)

[5] The observation here is in the spirit of Donald Davidson's point that an account of weakness of will must do justice to the fact that irrationality is a "failure within the house of reason" ("Paradoxes of Irrationality," in *Problems of Rationality*, Oxford: Oxford University Press, 2004, pp. 169–187). Both accounts of weakness of will that assimilate it to compulsion and accounts that appeal to a brute act of will place the phenomenon outside the house of reason. My point in the text is that if, as Wallace claims, akratic action reflects our capacity to "treat our disposition to do what we ought as a further desire from which we set ourselves apart, choosing to act in a way that is at variance with our reflective better judgment" (Wallace, "Normativity, Commitment, and Instrumental Reason," p. 10), this is because it reflects reason's capacity to undermine itself.

The possibility of passive agency

Two case studies

Even if we cannot wittingly defy our own reason with a brute "act of will," it does not follow that there is no more to the will than reason in its practical capacity. To the contrary, our will can operate independently when our reason is silent about what we should do. We can, for example, intentionally do something even when we take ourselves to have equally good reason to do something else. In such cases, we are the agents of our behavior. Nonetheless, there is an important sense in which we are passive in relation to this exercise of agency: if we have not ceased to identify with our capacity to reason, then to the extent that our actions are not responsive to the exercise of this capacity, *we* – we who are, essentially, rational beings – are *bystanders* to what we voluntarily do.[1]

Passive agency is, importantly, a form of *agency*. Accordingly, the passivity that characterizes such agency is distinct from the passivity of someone who fails to act because her behavior is the "wayward" – "deviant" – effect of her normative verdict (or of the intention to which this verdict gives rise).[2] The victim of wayward causation is a bystander to something – a nonagential movement of the body – that just happens to her. In contrast, what just happens to a passive agent is her action itself. Such an apparently paradoxical phenomenon is possible, I will try to show, because one need not exercise one's capacity to reason in order to determine one's will. This means that rational animals can act in the manner of nonrational animals, even as they continue to identify with their reason. And this means that there is an important sense in which rational animals can be alienated from – mere bystanders to – their own will.[3]

In the pages that follow, I will explore some of the forms that this self-alienation can take. I will begin with two rather extreme cases. Having thereby called attention

[1] The phenomenon to which I refer here satisfies the condition Frankfurt attributes to a "wanton" – an agent who "does not identify himself in a sufficiently decisive way with *any* of his ... first-order desires," and who is thus "remove[d] ... from his will so that his will operates without his participation" ("Freedom of the Will," p. 21).

[2] There is an extensive literature on this phenomenon – and on what distinguishes such action-preventing causation from the sort that (allegedly) suffices to generate an action. For two early papers, see Donald Davidson, "Freedom to Act," in *Actions, Reasons, and Causes* (Oxford: Oxford University Press, 1980, pp. 63–82), Harry Frankfurt, "The Problem of Action," in *The Importance of What We Care About* (New York: Cambridge University Press, 1988, pp. 69–78).

[3] Note that in identifying "willing" with "exercising one's agency," I part ways with those, like Kant,

to the metaphysical possibility of passive agency, I will then turn to some more everyday examples. I will conclude by explaining why passivity of this sort is essential to everything we rational beings intentionally do.

Bartleby the scrivener does many things in the long short story that bears his name. For a while he scribbles away at his desk, copying every legal document his employer places before him. But even at this early stage of his employment, something is not quite right as he works on "silently, palely, mechanically." Even as he is thoroughly absorbed in his task, Bartleby is also thoroughly detached.

The problem manifests itself in the fact that there are many things that Bartleby is not moved to do – including many things that any reasonable scrivener would do. As the story progresses, the list of such tasks grows ever longer. Soon Bartleby is spending each day standing in the corner of his employer's office, staring at a blank wall. Bartleby prefers not to work. He prefers not to sleep outside the office. Eventually, he prefers not to eat.

What sort of agent is Bartleby? I leave to one side the perversity of his preferences. The more significant feature of his agency is the status of these preferences, or – to put the same point somewhat differently – his relation to them and to the actions-cum-inactions to which they give rise. The problem is revealed whenever his employer tries to reason with him. "Bartleby," this mild-mannered gentleman desperately pleads, "These are your own copies we are about to examine. It is labor saving to you, because one examination will answer for your four papers. It is common usage. Every copyist is bound to help examine his copy. Is it not so? Will you not speak? Answer!"[4] And a few weeks later: "Are you ready to go on and write now? Are your eyes recovered? Could you copy a small paper for me this morning? Or help examine a few lines? Or step round to the post-office? In a word, will you do any thing at all, to give a coloring to your refusal to depart the premises?"[5] After several such failed attempts to elicit from his employee any sort of justification of his behavior, this remarkably patient man (patient, in large part, because he abhors confrontation) reaches the conclusion that Bartleby is simply "more a man of preferences than assumptions."[6] The point, I take it, is that Bartleby has no assumptions about the desirability or justifiability of his preferences. He does not assume that anything is worth doing. And so he cannot understand that anyone might think he has reason to act differently.

If Bartleby had thought that a compelling case could be made for his refusal to do his job, then he would have *assumed* something that could be challenged. In assuming that one's course of action is justified, one implicitly acknowledges that apparent challenges to this assumption are *relevant* to whether one really is justified. One is thus open to the intelligibility of changing one's mind. What is so frustrating – and so funny – and

---

according to whom the capacity to will depends on the capacity to reason.

[4] Herman Melville, *The Works of Herman Melville, the Piazza Tales,* Standard Edition, Volume X (New York: Russell & Russell, Inc., 1963), "Bartleby the Scrivener," pp. 19–65.

[5] Ibid., p. 51.

[6] Ibid., p. 49.

so "inhuman"[7] – about Bartleby is that he does not understand the game of giving and responding to reasons for action. This is what his "resistance to the doctrine of assumption" amounts to. When the narrator demands that he consider the folly of his ways, the only sort of response he has in his repertoire is a self-description: he does not prefer to do what his employer prefers him to do, and that is the end of the matter. In noting this evident fact, he might just as well be describing somebody else.

Insofar as Bartleby does not employ his reason to determine what is worth doing, he is like a nonrational animal. But unlike a nonrational animal, he gives no further indication of lacking the capacity to reason. This is what his employer indicates when he speaks of Bartleby's "inhumanity."[8] Cows and toads are not human beings. But for precisely this reason, cows and toads cannot be *inhuman*.

In effect, Bartleby is a case study in what someone would be like if Hume's most radical skepticism about the possibility of practical reasoning were correct. (Like the person whom Hume characterizes as preferring "his own lesser good,"[9] Bartleby "prefers not to be a little reasonable."[10]) Precisely because, like the agents of interest to Hume, Bartleby is *not* a nonrational being, his unresponsiveness to reasons renders him a passive bystander to his behavior. It would not be off the mark to describe him as "stubborn," and even "willful." But, as his employer well understands, this willfulness is nothing more nor less than a manifestation of his extreme "passivity,"[11] which is, in turn, a function of his extreme dissociation from the forces that move him to act.

Bartleby is a very perverse agent. Nonetheless, he appears to illustrate a genuine possibility: the possibility of being a passive bystander to one's own agency. Before I consider the importance of this possibility in the life of an ordinary, everyday rational being, I want to turn to a second, very different sort of, degenerate case.

The action that is at the center of Joseph Conrad's *Lord Jim* is as human as Bartleby's actions are alien. But the fact that Jim is the sort of agent who makes assumptions about what he has reason to do is inseparable from the fact that he – and the narrator Marlowe – have so much trouble making sense of what he does. This difficulty is especially instructive for my purposes. Though it does not support the Alleged Possibility, there is nonetheless an obvious sense in which this action is "contrary to Jim's best judgment."

Jim would rather die than jump. Yet he jumps rather than staying on board the ship that he believes is about to go down. How can this be? We have just seen that it is, in principle, possible for a rational being to be a passive bystander to the forces that move him to act. In Jim we have an example of what it is like to be a passive

[7] Ibid., p. 30. What he actually says is that there was not "anything ordinarily human" about Bartleby.

[8] Op. cit.

[9] David Hume, *A Treatise of Human Nature*, ed. by David Fate Norton and Mary J. Norton (Oxford: Oxford University Press, 2000), "Of the Passions," 2.3.3, page 267.

[10] Melville, *The Works of Herman*, p. 44.

[11] Ibid., p. 34.

bystander to one's action, even though one is not in the least bit indifferent to the considerations against acting this way.

Jim describes his shameful deed to Marlowe. He jumped, he confesses. Yet he also suggests that he was not really present at the event. "I had jumped ... 'He checked himself, averted his gaze ... 'It seems,' he added." "'Looks like it,' I muttered." "'I knew nothing about it till I looked up,' he explained hastily. And that's possible too. You had to listen to him as you would to a small boy in trouble. He didn't know. It had happened somehow. It would never happen again."[12]

In spite of himself, Jim jumped. That's a very natural way to describe what happened – what happened *to* him. Yet he did *jump*. He wasn't pushed. Nor did he trip overboard into the rowboat that would take him away from the spot where, he was convinced, something horrible was about to occur. So, in what did his passivity consist?

The answer is provided by Marlowe's psychologically acute description of the circumstances that lead up to the fateful transition from ship to rowboat. The ship had received a terrible wound. Anyone who knew anything about nautical matters would have been convinced that at any moment it would begin to sink fast. This fact alone would not have been enough to turn Jim into a passive agent. The key additional ingredient is the hundreds of human beings, fast asleep, unaware that they are about to suffer a horrible death. The thought of their terrified, pointless thrashings overwhelms Jim. It prompts in him a feeling of absolute hopelessness. He is convinced that there is absolutely nothing he can do to avert a calamity.

> He stood still looking at these recumbent bodies, a doomed man aware of his fate, surveying the silent company of the dead. They *were* dead! Nothing could save them! There were boats enough for half of them perhaps, but there was no time. No time! No time! It did not seem worth while to open his lips, to stir hand or foot. Before he could shout three words, or make three steps, he would be floundering in a sea whitened awfully by the desperate struggles of human beings, clamorous with the distress of cries for help. There was no help. He imagined what would happen perfectly; he went through it all motionless by the hatchway with the lamp in his hand – he went through it to the very last harrowing detail.[13]

He was not afraid of death perhaps, but I'll tell you what, he was afraid of the emergency. His confounded imagination had evoked for him all the horrors of panic, the trampling rush, the pitiful screams, boats swamped – all the appalling incidents of a disaster at sea he had ever heard of. He might have been resigned to die; but I suspect he wanted to die without added terrors, quietly, in a sort of peaceful trance. A certain readiness to perish is not so very rare, but it is seldom that you meet men whose souls, steeled in the impenetrable armour of resolution, are ready to fight a losing battle to

[12] Joseph Conrad, *Lord Jim* (Garden City, New York: Doubleday & Co., Inc., 1920, p. 81).
[13] Ibid., p. 62.

the last: the desire of peace waxes stronger as hope declines, till at last it conquers the very desire of life.[14]

Because, unlike Bartleby, Jim has very strong opinions about how he ought to behave, he can succeed in disengaging his action from his reason only by going into a sort of trance. He achieves this state by giving up all hope of being able to put his reason to any useful service. The "peace" that results is the peace of the dead.

Overcome by the mutually reinforcing feelings of hopelessness and helplessness, Jim dissociates himself from his capacity to reason. He ceases to even consider the possibility that he might put this capacity to good use. In giving up on his reason, he dissociates himself from his actions. They become for him mere things that happen. He becomes a bystander to these happenings – an extremely *absent-minded* bystander. When he discovers that he has jumped into the rowboat, he is like someone who discovers that he has opened the refrigerator door in his sleep.

Jim's jump is the action of a sleepwalker. Except, of course, that he is fully awake. It is his *reason* that is asleep at the wheel. His reason is sleeping during his fateful action because, by that time, he has abandoned the task of setting goals for himself. Because he no longer believes in his own power to make a difference, he lacks any motive for monitoring his activities. The dissociation from his agency that results is reinforced by his powerful desire to avoid thinking about what is likely to happen, no matter what he does.[15]

Other cases of passive agency

The contrast between Bartleby and Jim can perhaps be sharpened by comparing two other cases – altogether lacking in the bizarre and the tragic. As in the cases of Bartleby and Jim, in these cases, the fact that the agents are not putting their reason to practical use is tied to the fact that they are very imperfectly responsive to reasons. I hope to show, however, that this connection is a contingent one. Indeed, in considering these cases, I will be taking the first step toward stressing the extent to which the will's disengagement from reason is a characteristic feature of a well-functioning agent.

Imagine that someone is at a party, sitting next to a big bowl of potato chips. In the first case, the partygoer feeds herself one potato chip after another, blithely unaware

[14] Ibid., p. 64.

[15] For the purposes of this discussion, I do not want to assume that intentional action requires a special epistemic relation to what one is doing (e.g., "practical," nonobservational knowledge). For all I say about this case or the others that follow, we can do things for reasons without being aware of these reasons, or even of the fact that we are responding to them, or even that we are doing anything at all. So, too, the absence of such awareness is not essential to passive agency. As should be clear from the discussion that follows, someone can be a passive bystander to her actions even if she is aware of what she is doing, and even if – unlike Bartleby – she *assumes* that some actions are more justified than others. (For the locus classicus of the view that intentional action essentially involves a special sort of knowledge of what one is doing, see G. E. M. Anscombe, *Intention*, New York: Blackwell, 1958.)

of what she is doing as she animatedly discusses weakness of will with the person sitting on the other side of the table. In the second case, the partygoer watches her hand as it repeatedly dives into the bowl for another potato chip. All the while, she is eavesdropping on an extremely interesting conversation. But this does not prevent her from speculating, with mild amusement, about whether her feeding frenzy is likely to stop before she has emptied the bowl.

The first potato chip eater is like Jim in being dissociated from what she is doing because her mind is "absent." This is, it seems, importantly different from the case of someone who does not think about what she is doing, but is – as we say – fully present in her actions. If the best way to understand this "presence" is to regard the agent as responding to reasons in the way that any nonrational animal does, then, by stipulation, she does not occupy a point of view that renders her a bystander to her own actions. If, however, as I believe, she is best understood as (unconsciously) endorsing the fact that she is responding like a well-functioning nonrational animal, then she, too, is passive – without being absent.

I will work my way up to the more common forms of this passivity by starting with the second potato chip eater. This agent is fully aware of what she is doing. Let's stipulate, too, that she is capable of halting the movement of her hand whenever she thinks this would be a good thing to do. It's just that, like Bartleby, she never forms any such opinion. Now, this could be because, like Bartleby, she is not capable of putting her reason to use in this way. Let us suppose, however, that she has no such congenital inability; she is simply *unwilling* to consider the reasons she may have for or against eating all those chips.

Not only is it possible for someone to be indifferent to the desirability of her actions, someone can also take an interest in the desirability of her actions without having any interest in being constrained by these evaluations. Kierkegaard's aesthetes are an extreme example of agents who relate to their actions in this detached way. As Kierkegaard explains, a true aesthete is someone who has chosen to relate to every event and state of affairs – including every aspect of her own inner life – as if it imposed no constraints on her choices. From the point of view of a pure aesthete, the only significance of eating potato chips and of the desire to eat them is their *aesthetic* significance as prompts to the imagination. The aesthete allows things to happen, in her and to her, moved only by the desire to find interest in the passing show. "Given," she thinks, "that I am moved to eat potato chips, what interesting story can I tell myself about this desire and the effects it causes in me and the world? How many variations on this story am I capable of imagining?"[16] No ordinary human being could sustain this posture for very long. But if Bartleby is a metaphysical possibility, then this sort of self-alienated agent is a metaphysical possibility too.

[16] For a brief summary of the type, see Soren Kierkegaard, "The Rotation Method," in *Either/Or*, Vol. I, trans by David F. Swanson and Lillian Marvin Swenson (Princeton, NJ: Princeton University Press, 1959, pp. 280–296).

More importantly, one need not be an aesthete in order to intend/choose/decide to be a passive bystander to one's own agency on any given occasion.[17] As I suggested in my brief comment about the agent who is "present" in her actions without being guided by her reason, a rational being can form opinions about the desirability of what she is doing without doing so in order to constrain her will. It is also possible for someone to endorse what she does even though this action is not responsive to reasons. Some potato chip eating of the second sort is like this. So, too, are some cases in which a person gives her emotions free rein.[18] When someone stamps her foot or brings her fist crashing down on the table, this need not be something she does as a means to an end – not even the end of expressing her emotion. Yet her emotion would not have moved her in the way that it does had she not assumed that she lacked sufficient reason to prevent it from doing so. Her reason is, in other words, the sort of bystander who is capable of intervening.

Something similar is true of much habitual action. We rightly believe that in many circumstances it is better to allow various well-conditioned dispositions to exert their characteristic effects on our bodies, unhindered by any contemporaneous considera-

[17] The position I lay out in this section can be contrasted with the position Kieran Setiya defends in *Reasons Without Rationalism* and "Sympathy for the Devil." According to Setiya, "one can act for a reason without believing that there is a reason that counts in favor of what one is doing, or any respect in which it is a good thing to do, and without either of these propositions seeming to be true." ("Sympathy for the Devil," p. 92) "One need only believe that one is doing X because P, so long as there is the right sort of connection between one's action in doing X and this belief... . Since the right connection need not involve the belief that one's reason for doing X is a good reason for doing it, one need not have that further belief."(91) Though this is not the place to offer a thorough response to this position, it should be clear what sort of response I favor: leaving aside the cases in which one distances oneself from the point of view of one's earlier self, either Setiya is simply calling attention to the possibility of passive agency, or he is suggesting that passive agency is the paradigm case of acting for a reason. In the first interpretation, he is not challenging the view that whenever we put our reason to practical use (in order to determine the will), we act under the guise of the good. But given that he takes himself to be offering just such a challenge, the problem with his account is simply that he treats aberrational cases as paradigm cases. If I do not endorse the grounds for doing X which I cite to explain why I am doing X, then I am a mere spectator to my own behavior. Of course, when asked why I am doing X (as opposed to why I *did* X – again, everyone is a bystander to her past self), I can offer an explanation without thereby ceasing to take my action to be justified, or suggesting that I do. Indeed, the person who asked me why I am doing X will assume that in reporting the considerations that move me to act, I take these considerations to justify my action, and that I would not have acted this way had this not been the case. This is why, she will rightly assume, I am able to explain my action by appealing to the fact that I am treating these considerations as justifying reasons. Even if, as Setiya suggests, it is consistent to believe that (i) one is acting as if F is a sufficient justifying reason to X and that (ii) F is not really a sufficient justifying reason to X, holding these beliefs simultaneously is a paradigm case of self-alienated agency. This is, again, precisely because it is a case in which one's belief about what one has reason to do is dissociated from any interest one has in determining one's will.

[18] For a discussion of so-called "arational" actions, see Rosalind Hursthouse, "Arational Action," *The Journal of Philosophy* 88, no. 2 (February, 1991): 57–68. For a discussion of such actions that is in sympathy with what I say here, see Joseph Raz, "Agency, Reason, and the Good," in *Engaging Reason: On the Theory and Value of Action* (Oxford: Oxford University Press, 1999, pp. 12–45, 36–44).

tions (conscious or unconscious) regarding what it makes sense to do. It is worth stressing, moreover, that this need not involve reverting to some set of rigid routines: habitual responses can be extremely complex and subtle; they can require exercising sophisticated skills.[19] When such habits "take over," the agent will usually be doing what she has sufficient (and even decisive) reason to do (using her backhand, repeating a melody with a slight modification, adjusting her tone of voice, making room for someone to pass by, putting her shirt on right side out with the collar open in the front, etc.). What's more, she will be doing these things *because* she has sufficient reason to respond to her circumstances in this way. Often such actions resemble Jim's jump insofar as the agent is not aware of what she is doing. But, like Bartleby, someone can also be an *observant* bystander to her own habitual behavior. So, too, she can be "present" – or as we sometimes put it, "absorbed" – in what she is doing.

The behavior associated with *bad* habits is unresponsive to the decisive reasons to *refrain* from behaving this way. Often – usually – this is because the impulse to behave this way has motivated the agent to employ her reason in its favor – at least for the few seconds that are needed to nibble a fingernail. As long, however, as someone's *reason* is truly a bystander to what she does, it is possible for her to be a *critical* bystander. For, as we have seen, it is possible to disapprove of an action without being opposed to acting this way. Though, unlike Bartleby, such a self-critical agent believes she has reason to act otherwise, she remains dissociated from what she is doing because she has no interest in determining her will. To be sure, someone cannot be *indifferent* to how she acts while *taking a stand* on how she has reason to act. The point is that to take *an interest in a practical matter* is not necessarily to take a *practical interest in the matter.* Our reason need not speak to us in our capacity as agents, even if it speaks to us about actions, and even if it speaks to us about how *we* have reason to act.[20]

*

[19] For more on this point, see Peter Railton, "Practical Competence and Fluent Agency," in *Reasons for Action*, ed by David Sobel and Steven Wall (Cambridge: Cambridge University Press, 2009, pp. 81–115).

[20] For an account of intentional action according to which all rational agents occupy an essentially theoretical stance toward their behavior, see J. David Velleman *Practical Reflection* (Princeton, NJ: Princeton University Press, 1989) and *How We Get Along* (Cambridge: Cambridge University Press, 2009). Note that in Velleman's account, the agent's interest is "theoretical" in an additional sense: it is, essentially, an interest in being able to *explain* what she does in terms of the features of her psychology. It is also worth stressing, however, that Velleman takes this interest to be motivating: the desire to understand what we are up to moves us to behave in a way that we can understand. In *How We Get Along*, he acknowledges that in addition to this action-guiding desire to understand our behavior, we can take the sort of detached stance that interests me here. As he explains, "the role of backstage onlooker requires a degree of mental compartmentalization, since it requires a skeptical suspension of the very self-interpretation that is being applied and enacted in the other two roles. This role must therefore be carried out in a partially encapsulated mental process, a train of thought selectively insulated from the mental processes involved in self-enactment and first-order self-interpretation. This train of thought contains interpretations of the agent's behavior that will not be enacted and are not meant to be. It

In the chapter's first main section, I, in effect, argued that, *provided one employs one's reason in order to determine one's will*, intentional action is "under the guise of the good" (it is under the guise of good-enough reasons). In this section, I have argued that, *if one does* not *employ one's reason in order to determine one's will*, then one will not be guided by the resulting normative judgments. In my account, then, the "judgment externalists" are right to insist that someone can believe she has overriding reason to do X without being moved to do X.[21] Yet the judgment "internalists" are right to insist that there is an important sense in which it is not a contingent matter whether someone's contemporaneous normative judgment moves her to act: as long as this judgment expresses her attempt to determine what to do, its effect on her behavior does not depend on anything else about her motivational profile; it determines her will immediately, that is, without the mediation of any further representation.

A given judgment either is or is not the product of reasoning that is practical; the aim of determining the will either is or is not internal to the inferences that are

---

is therefore detached from the feedback loop of interpretation and enactment that gives the agent his experience of selfhood in the practical realm. The result is that, from the agent's point of view, this process of self-observation seems to take place in a consciousness that is his own and yet separate, watching from behind the scenes as he plays himself" (David Velleman, *How We Get Along* (Cambridge: Cambridge University Press, 2009, p. 93)).

[21] Here I am in agreement with Kate Manne, whose paper on this issue I discovered only after completing this one. (See "Tempered Internalism and the Participatory Stance," in *Motivational Internalism* ed. by Gunnar Bjornsson, Caj Strandberg, Ragnar Francen Olinder, John Eriksson, and Fredrik Bjorklund, Oxford: Oxford University Press, 2015, pp. 260–281)). In trying to make sense of the possibilities to which externalists about *moral* judgments appeal, Manne suggests that these are cases in which "the agent's sense of *agency* has somehow gone missing" (263). "A moral judgment made by an agent about what she herself (currently) ought to do," Manne explains, "will entail motivations on her part to act in accordance with this judgment, provided that she takes the practical stance toward herself which is fitting for such judgments" – that is, provided that she takes up "the participatory stance."(263) Though I endorse Manne's basic point, I take issue with some of the things that she says. In particular, (i) it is misleading, at best, to describe someone who is alienated from her agency in the relevant respect as "seeing [herself] as passive rather than active"(267). The point is that she sees herself as passive in relation to her own agency. This is why, (ii) contrary to Manne's claim, such agents are right to "credit these bits of behavior to them[selves] as being their actions"(268). It is why Jim is right to regard his jump as "expressive of [his] agency for which [he] will have to answer"(268). (iii) I also part ways with Manne insofar as she suggests that it is possible to maintain the participatory stance in relation to one's own will without maintaining it in making one's moral judgments. Or rather, it seems to me that this would be possible only if these judgments were not judgments about what one has reason to do. But then it would be this feature itself that would explain the lack of motivation. The appeal to the agent's failure to take up the participatory stance would be superfluous. As my comments about good habits suggest, (iv) it is also a mistake to claim that this sort of "dissociation" is necessarily defective. Finally, (v) I think it is important to distinguish the subjects in the Milgram experiments from the other cases Manne describes. These agents are struggling to overcome what they experience as a temptation to do what they are committed to not doing. This is surely, as Hilary Bok argues ("Acting Without Choosing," *Nous* 30, no. 2 [June, 1996]: 174–196), because they fail to appreciate that their resistance to continuing to administer the shocks is a reason to reconsider their decision to participate in the experiment. The point is that their agitation is a symptom of the fact that they are not passive bystanders to their own agency.

constitutive of an episode of reasoning. If a person's all-things-considered normative judgment is the product of reasoning with a built-in practical aim, then – as long as she does not abandon this aim – this judgment cannot fail to constrain her will. If, on the other hand, her all-things-considered normative judgment lacks this etiology, then it cannot impose any such constraint. Again, this is why both judgment internalists and judgment externalists are mistaken to identify the possibility of being unmoved by one's reason with the possibility of practical irrationality.[22]

I have already noted one respect in which, despite the impossibility of clear-eyed akrasia, our actions can be "contrary to our best judgment" (viz., they can be the product of rationalizations, where this involves treating certain facts as sufficient reasons for action, even while being aware that this reasoning is unlikely to survive careful scrutiny). My reflections on passive agency suggest another phenomenon that can accurately be so described. Though not even depression or a strong temptation can push or pull someone to do something that, at the moment of action, she believes she would rather not do, all things considered, such conditions can lead someone to take a detached attitude toward her own agency. They can prompt her to cease to care about whether she does what she deems she has most reason to do. They can move her to retreat to the position of a mere bystander. In short, when "giving in to temptation" does not take the form of the sort of bad reasoning we call "rationalizing," it involves failing to be constrained by one's own normative verdicts under circumstances in which these verdicts are dissociated from any interest one may have in determining how to act.

*

[22] It is not, as James Dreier puts it (speaking for most others), that a "connection failure [between judgment and motivation] is not impossible, but irrational" ("Can Reasons Fundamentalism Answer the Normative Question?" In *Motivational Internalism*, ed by Gunnar Bjornsson, Caj Strandberg, Ragnar Francen Olinder, John Eriksson, and Fredrik Bjorklund (Oxford: Oxford University Press, 2015, pp. 167–181). Rather, such a connection failure is possible, but not irrational (though, of course, it may well be unresponsive to reasons). Again, precisely insofar as practical deliberation "takes place with an eye to acting directly on the conclusion of that deliberation" (Sigrun Svarvasdotir, "How Do Moral Judgments Motivate?" In *Contemporary Debates in Moral Theory*, ed. by James Dreier (Malden, MA: Blackwell Publishing, 2006, pp. 163–181)), as long as this commitment remains, the conclusions of practical deliberation cannot possibly "have no resonance in [the deliberator's] motivational system" (175), and so there can be no form of irrationality that consists in this failure of resonance.

The desirability and necessity of passive agency

Purely expressive actions and actions that manifest bad habits are not the only actions governed by mechanisms that are unresponsive to reasons. Indeed, we rely on some such mechanism whenever we act. This is because for every action, there are alternative actions which, from our point of view, are equally desirable. We can describe any action in enough detail that either (i) we have no reason to prefer this action to any others that can also be so described or (ii) we have good reason not to take the time and effort necessary to discover any such reason. Think, for example, of the indefinite (infinite?) number of ways in which someone can "walk up the stairs." Whichever way she plants her left foot, on whichever spot, she is doing "this" intentionally insofar as "this" is "going upstairs." It goes without saying (or thinking) that no reasonable person would be concerned to ensure that she was responding to reasons for planting her foot *this way rather than that way*. Accordingly, both she and we can take it for granted that a reasonable person would endorse this measure of arbitrariness in her ascent up the stairs.

Of course, to characterize this as the stance of a "reasonable person" is not to rule out the possibility that someone could care about even the most minute differences among the indefinite (infinite?) number of ways in which she might walk up the stairs. But even if a being with such concerns were to have the epistemic powers (and the stamina!) necessary to be guided by these concerns, she could not completely forego the role of passive bystander in relation to her own actions. For not even such a hyper-discriminating being has the power to alter the fact that every action is a particular event.

Without the capacity to reason, we could not understand the distinction between act-types and act-tokens. Yet our reason can do nothing to help us select an act-token *as such*. This means that the conditions of agency impose limits on the extent to which our actions can be responsive to reasons. For in order to do anything at all, we must will to perform this, here, particular act.[1]

To respond to reasons is, necessarily, to respond to one's circumstances in ways that are not arbitrary. But if no reason can be offered for doing *this*, here, now, rather than *something else*, here, now, then doing *this*, here, now is, to this extent, arbitrary.

[1] I recently discovered a passage in which Jonathan Dancy makes this point. "A reason," he says, "is never a reason for a particular act; it is and can only be a reason to act in a certain way. So the fact

What follows is not, as Christine Korsgaard suggests, that "particularistic willing" is impossible.[2] Rather, because all actions are particulars, a rational being is necessarily passive in relation to some aspect of her will.

This brings us back to the cow and the toad. Because these animals cannot take themselves to have reasons for doing one thing rather than another, nothing they do is arbitrary from their own point of view. This is another way of saying that they cannot be passive in relation to their own actions. As we have seen, however, things change when the capacity to reason is added to the capacity to will. Unless a rational being ceases to identify with her reason, she cannot manifest a pure exercise of will without being passive in relation to her own action. If one is a cow or a toad, then, from one's own point of view, the most minute aspects of one's actions – the exact placement of one's mouth on the grass, the exact angle of one's hop – are no more arbitrary than the most general features of these actions. If, however, one is a rational being, then, from one's own point of view, the purity of one's will is inseparable from its arbitrariness. This means that if the will of a rational being is always to some extent unconstrained by her reason, then to just this extent, a rational being is always a passive bystander to her own actions.

*

I have argued that the capacity to reason is the *in*capacity wittingly to defy one's own reason. This means that there are limits to the possible ways in which we can be passive in relation to our own actions – limits imposed by the necessary conditions of agency. But to have the capacity to put one's reason to practical use is not necessarily to exercise this capacity on any given occasion. This means, I have also argued, that it is possible for us to be passive in relation to our own actions. Sometimes, this possibility gets us into trouble. Think, for example, of poor Jim. But far more often, it saves us from paralysis, and clumsiness, and exhaustion. It preserves our sanity. Not only, moreover, is it often desirable for us to be bystanders to our own actions; some measure of passivity is also a necessary condition of everything that we do. This is because the capacity to reason is the capacity to distinguish what is arbitrary from what is not and because we cannot draw this distinction in a way that eliminates every element of arbitrariness from our actions. Even as the necessary conditions of agency ensure that a person's will cannot defy the verdicts of her reason, so too, the necessary conditions of agency ensure that, without the help of the will, a person's reason is powerless to determine what she will do.

---

that I owe you five pounds is a reason for me to pay you back when you ask for the money, but it is not a consideration that favours any particular way of paying you back; it leaves many aspects of that recommended reimbursement unspecified." (Jonathan Dancy, *Practical Shape: A Theory of Practical Reasoning*, New York: Oxford University Press, 2018, p. 31). Dancy credits Prichard with the basic insight.

[2] See Christine Korsgaard, "Practical Reason and the Unity of the Will," in *Self-Constitution:*

Agency, Identity, and Integrity (Oxford: Oxford University Press, 2009, pp. 59–80).

14. Three for the price of two

Jonathan Dancy

In this chapter, I outline the way in which, moving from two views already formed, I came to a third without further effort. The two together gave me the third more or less for free. Such luck is rare in philosophy.

1 The first view: holism in the theory of reasons

In early papers (Dancy 1981, 1983) and my first book on ethics (1993), I proposed and developed the then heretical claim that a consideration that is a reason in one case need not be the same reason in another. This idea came to me while thinking about W.D. Ross's theory of prima facie duty; I just could not see why my having promised to do it should *always* give me (as I would put it) some reason to do it. Since the idea that what is a reason in one case is always a reason in any other was held at the time to be analytically true, nobody had thought to defend it. Once questioned, it effectively collapsed (I would claim).

The main interest of this was, of course, that it bid fair to undermine principled conceptions of ethics. A moral principle is a statement that records the invariant relevance of a consideration, whether that consideration be enough to make an action a duty proper or merely a prima facie duty. But this relevance to ethics is not the main point for present purposes. More important is what came with it. For if a reason in one case may be no reason in another, and even be a reason on the other side, there would have to be some explanation of why there is that change of relevance. And that explanation would most probably point to some other consideration, absent in the first case but present in the second and capable when it occurs of preventing the first consideration from being a reason. Call this second consideration a disabler. Equally, there can be enablers, which turn considerations that in other cases were not reasons or would otherwise not have been reasons into reasons in a new case. So we have a distinction between enablers and disablers to add to the notion of a reason. And in addition to that, there is a distinction between what I called intensifiers and attenuators, that is, between considerations that make what is already a reason into a stronger or weaker reason than it otherwise could have been.

This yielded a noticeably more extensive toolkit for understanding the ways in which considerations can combine to make an action a duty – or more simply what one has most reason to do. It amounted to a more complex theory of practical relevance – that is, of the sorts of relevance a consideration can have to one's practical decisions. Relevance is not just for or against; there is also the relevance of an enabler, a disabler, an attenuator and an intensifier. And for all I knew, there might be further forms of relevance. John McDowell had written of silencers – one consideration silencing another as a reason; Joseph Raz introduced the notion of an exclusionary reason – a reason that excludes some other reason from consideration (McDowell 1998, 17–18, 55–56,

90–93; Raz 1975). It seemed that the study of practical relevance was in its infancy. But at least we were beginning to unfold more of the complexity of ordinary practical thought. And of course the same distinctions are available in epistemology.

2 The second view: realism in the theory of motivating reasons

For nearly forty years after the publication of Donald Davidson's 'Action, Reasons and Causes' (1963), the view that he there propounded was philosophical orthodoxy. This view, which I call psychologism in the theory of motivating reasons, was that the reasons for which we act are our own psychological states, such as our beliefs, intentions and desires. To act for a reason is to be caused so to act by a suitable combination of such states. You cannot be caused to act by the psychological states of others – at least not directly. And if you are caused to act by something other than a state of your own, you are not acting for a reason.

Against this view, I argued in my *Practical Reality* (2000) that, since most of the reasons for us to act in one way rather than another are such things as that it is a fine sunny day and that we have no more money (which are good reasons but not psychological states of ourselves), Davidson's psychologism makes it impossible for us to act for a good reason. Against that psychologism, I argued for what I called realism in the theory of motivating reasons, which amounted to little more than the idea that it must be possible to act for a good reason and that many things that are good reasons for us are not our own psychological states. They might, for one thing, be psychological states of others. But they don't have to be psychological states at all, without that preventing us from acting for them. Any state of affairs is the right *sort of thing* to be a good reason.

I also argued against a rather different claim, that the reasons for which we act are propositions with a certain content. The form of the claim 'his reason was that p' tempts the unwary to think that his reason was 'that p' – where 'that p' represents a proposition. I argued against this that no proposition – not even a true one – is capable of being a reason, of counting in favour of, or against, acting in one way rather than another. And if they cannot count in favour of action, they cannot be the reasons for which we act, on pain again of making it impossible for us to act for a good reason.

I called my view 'realism in the theory of motivating reasons'. It helped that nobody ever denied what one might call 'realism in the theory of normative reasons': that only what is the case can be a good reason for doing one thing rather than another. But one can act for a reason that is not the case, sadly. There is no inconsistency in saying that the reason for which he did it was that it would increase his pension, something about which he was sorely mistaken.

3 The favouring relation

We need to add to these views something much less tendentious, and not in any way peculiar to me, an account of the favouring relation. A consideration that is a reason for action (or for anything else such as belief or intention) is one that favours so acting (or so responding, more generally). This is a truism, but it is an important one because it places the favouring relation at the centre of one's understanding of reasons. We have favouring vs. disfavouring, enabling vs. disabling and intensifying vs. attenuating: three distinctions (at least). But it is the favouring relation that is central – the 'being for' side of the 'for vs. against' distinction.

Can we give any account of this central relation? Well, perhaps all we can do is to specify its form. I understand the favouring relation as a three-place relation that ties together a consideration, a response (such as a way of acting or a believing) and an agent from whom that response is called for. The same reason may be a reason for you to act in one way and for me to act in another. Our actions will be our responses to that reason. Equally, the same consideration may be a reason for me to form the intention to act and for you to believe that I will form that intention. Here my response to that reason is the forming of an intention, while your response is the forming of a belief. Belief is as much a response to reasons as is acting, or intending. So what is favoured is always a response of some sort. What does the favouring is always a matter of fact, or a feature of the situation. And then, since the same consideration may be a reason for me to start running and for you to hide, we need a place in the favouring relation for the relevant agent. So an instance of the favouring relation will have three elements: reason, response and agent.

4 Aristotle on practical reason

With these preliminaries, we can now move to the payoff. Aristotle is supposed to have held that, in addition to theoretical reasoning which takes us from reasons to belief, there is such a thing as practical reasoning, which takes us in just the same way from reasons to action. The *locus classicus* for this is his *Ethica Nicomachaea* 7.3, 1147a26–31, where he writes

> whenever some one thing is derived from them [a pair of premises, one universal and one particular], that conclusion must in the one case be asserted by the soul, and in the case of practical reasoning immediately be done; e.g. if everything sweet should be tasted, and this is sweet (which is one of the particular premises), the agent who is able and is not held back must simultaneously actually do this'.[This is my translation. Note the 'immediately' and 'simultaneously' here. These are different words in Greek, but they both refer to time.]

Aristotle cast this claim in terms of his own discovery, the syllogism. There is the theoretical syllogism, which comes in various forms of which this is one: all cows eat grass; this is a cow; so this eats grass. The conclusion of this reasoning is, Aristotle says, a belief or acceptance. And there is the practical syllogism, which seems to be of this sort of form: dry food is good for a man; chicken is dry food; I am a man; this is chicken; so – I eat. Here the conclusion is an action, my eating the chicken before me.

Note that the action is not held here to be a sort of secondary conclusion, preceded by the primary conclusion that I ought to each this chicken. Nor does Aristotle seem to think that one should first form an intention to act – the intention being the primary conclusion. No: there are such conclusions, and one can draw them from the same reasons, but Aristotle's view is that one can go straight from those reasons to the action they favour, without passing through those other possible conclusions on the way.

This position of Aristotle's is widely held to be indefensible, with some reason. One way of putting the point is that eating a chicken cannot be the conclusion of an inference; the most one could infer from those premises is something like 'I should eat that chicken', or perhaps 'I will eat that chicken'. Action, it seemed, may happen as a result, but it cannot itself stand in the same sort of relation to the considerations adduced in the reasoning that belief can stand. Aristotle's view is impossible.

But now: one's action is as much a response to one's situation as any belief can be. Actions can be favoured by aspects of one's situation – by considerations adduced –

just as believing and intending can be. And actions can be done, and beliefs formed, in the light of those considerations. So why should we not say that actions can stand more generally in just the same sorts of unmediated relation to considerations adduced as belief and intention can stand? And would that not be enough to reinstate a conception of practical reasoning, one which would however not need to be syllogistic?

It seemed to me therefore that my two views might lead to a neo-Aristotelian conception of practical reasoning. This was the point at which a third view emerged in the wake of the first two. Of course it is not a straightforward consequence of the first two; one cannot *derive* it from them. What they do instead is to put us in a position in which the possibility of the third view becomes visible to us.

5 The third view

So the idea (which I developed in detail in my *Practical Shape* (2018) was that we can give a general account of reasoning under which reasoners reason from things they accept to a response of some sort, that response being the one most favoured by the considerations they have adduced. If the reasoning is 'theoretical', the direct response will be a belief; if the reasoning is 'practical', the direct response will be an action. (This is of course just terminology.) There is also an intermediate form of response, the forming of an intention, which is still practical in a good sense. (I call this 'intermediate' not because it always comes between belief and action – it doesn't – but because it is both belief-like, since it has a content, and action-like, since it is a response to practical reasons.) But I will say little here about this third form of response, for lack of space. I restrict myself to the relation between practical and theoretical reasoning, reasoning to action and to belief.

Actions can be done either for a simple reason or in the light of more complex constructions of reasons. When I get off the bus because it is my stop, I am acting for a reason; but I would not count this as an action done in the light of reasoning. I reserve the title of 'reasoning' for more complex cases, where there are various considerations adduced, of varying relevance. The distinctions I introduced in section 1 give us ways of understanding how various considerations can combine to make a case for action, even when not all of them are independent reasons for acting in that way. Reasoning occurs when one puts together considerations which are relevant in their different ways, perhaps, and responds appropriately – that is, in the way most favoured by those considerations, taken as a whole. The reasoning is practical if that response is an action and theoretical if that response is a belief. And the practical response does not need to be built on the back of a theoretical one; it can be direct and unmediated.

There will be intermediate cases in which there are several reasons at issue, some on one side and some on the other, and the only question is which team has the most weight. Here we do not find all the complexity that can arise when my other distinctions (intensifying vs. attenuating, enabling vs. disabling etc.) come into play. One might therefore ask, of those intermediate and simpler cases, whether they are to count as reasoning. My answer to this is that it does not matter which way we go. As we will see, on my account acting for a reason is not radically dissimilar from acting in the light of reasoning; they differ only in the complexity of that to which one is responding in action.

One significant similarity between practical and theoretical reasoning, then, is that my distinctions between enabling and disabling and between attenuating and intensi-

fying apply with equal ease to both types of reasoning. Notice that I say 'types' here, rather than forms, because on my view all reasonings have the same basic form: there are considerations adduced, of varying relevance, and there is the response we make in the light of those considerations, taken as a whole. The 'varying relevance' here just means that some considerations will be independent reasons, while others will play other, secondary roles; they count as secondary because their roles are understood in terms of the basic notion, that of favouring, or of being a reason.

6 An example of theoretical reasoning

Here is a theoretical example to which we will apply these distinctions between forms of relevance to show how it can be done. You are a detective investigating the murder of the appallingly proud and domineering Lady Snobgrass. She was shot, and you find the gun in the butler's cupboard. Here, then, are some relevant considerations:

1. The gun was found in the butler's cupboard
2. There were no signs of its being a plant (that is, to have been planted there for you to find, so as to cast suspicion on the butler).
3. Your investigation has been conducted with great care.
4. There are no other serious suspects.
5. The butler had something (but not much) of a motive.

Let us consider the respective roles of these 'premises' and start by allowing that the first premise, considered alone, is an independent reason to believe that the butler did the deed. (What the 'considered alone' actually means here is a source of considerable difficulty, but I won't open that can of worms here.) That is, it favours believing that the butler did it. The second premise is more interesting. It seems to be relevant, but in what way? Perhaps it intensifies the reason given us by the first premise. But consider how things would have been if the second premise had been false. In that case, the first premise would have favoured believing that the butler was innocent. That is, the falsehood of the second premise would have turned the first premise from a reason to believe the butler to be guilty to a reason to believe him innocent. I take this fact to be part of the relevance of the second premise as it stands. For each premise, we need to consider how things would have been had that premise been false. (Perhaps this is a general rule for those trying to understand how reasoning works.) The first premise is such that if it has been false, we would have had one fewer reason to believe the butler to be guilty. The second premise is such that if it had been false, we would have had one fewer reason to believe the butler guilty and a very good reason to believe him innocent.

Of course, the way I have formulated the second premise is a bit cagey, since it allows for the later realisation that the gun was carefully arranged in that way by the butler

himself, in an attempt to divert suspicion. But such possibilities are an unavoidable feature of all reasoning of this sort.

What about the third premise? Here the suggestion I want to make is that it is effectively a comment on the reliability of the premises as a whole, though not of the reasoning since it concerns the input to that reasoning rather than what you do with that input once you have got it. So one thing it does is to support the idea that there is not much that you have missed, and this goes beyond any suggestion that the truth of your premises is comparatively solid.

As for the fourth premise, this seems to me not to be an independent favourer. It would favour believing that the butler did it only if the case against him were strong enough on its own, which it might not be. In a way it is more like a counsel of despair, or perhaps a note of caution. Now we know that we ought not to believe that the butler did it, even if he is the only serious suspect, unless the case against him can be made out sufficiently well. But this is an 'ought', or rather an 'ought not', and tells us little about what reasons we have, where reasons are understood as considerations favouring a certain response. We can perfectly well have some reason to believe the butler did it but not enough to justify coming to that conclusion. There is no principle of detection, or of other enquiry, that one ought to believe whatever one has most reason to believe, if one would do better not to believe anything at all yet.

In addition, the fourth premise depends for its significance on whether we have eliminated all other potential suspects, or whether we just haven't looked very hard.

Finally, the fifth premise. I take this to be an enabler. A strong motive would probably count as a favourer, but a motive of this weak sort seems to do little more than allow the butler to remain under suspicion. It hardly promotes the case for his guilt beyond that.

What is the conclusion from all this? The general point is that theoretical reasoning is very similar to practical reasoning in all the sorts of ways that I have outlined. We could indeed hope to map a detective's reasoning using the tools that have emerged. So to this extent, theoretical and practical reasoning are on a par.

7 Explaining reasons and reasoning

If practical reasoning is so similar to theoretical reasoning, what differences might there yet be between them? My suggestion here is that they differ mainly in the way we explain the ability of the considerations adduced to favour the relevant response. And this difference can be most easily seen when we ask, more simply and more generally, what explains the ability of the considerations that are reasons to be reasons. I suspect that practical reasons do differ from theoretical reasons on this front (though it is not important to my main thesis that they should). Joseph Raz argues (see his 2011, ch. 3) that practical reasons are explained by some relation to concerns or values, while theoretical reasons are explained in a different way. The standard account of that different way (not Raz's, though) is that theoretical reasons raise the probability of their 'conclusion' – the belief they favour. If so, they differ from practical reasons because the latter do not raise the probability of anything – or rather, it is not because they raise the probability of something that they are reasons. (For contrary views on this last point, see Thomson 2008; Kearns and Star 2009.)

Now all this might be wrong; such issues are surprisingly little discussed. Perhaps practical reasons raise the probability of a 'practical belief', such as the belief that one ought to act in such and such a way; this would hardly be surprising – but I would say that they raise the probability of that belief because they favour the action and not the other way round. (This is what I call in my *Practical Shape* the 'Primacy of the Practical'.) Perhaps some theoretical reasons make their conclusion more plausible rather than more probable; it may be that in philosophy, what we are trying to do – what I am now trying to do – is to achieve plausibility rather than probable truth. Perhaps all reasons are explained by relation to values: the value of truth, or the value of probability, or other more practical values.

8 Formally valid reasoning

In addition to reasoning that seeks to determine what conclusion is made most probable by the considerations adduced, there is also formal reasoning. Formal reasoning aims to guarantee the truth of its conclusion and to do so by virtue of its form. Take a standard case: simple *modus ponens*. The premises of any inference of this form, if true, guarantee the truth of the conclusion, *because of their form;* any other inference of the same form would be as effective. This is quite unlike ordinary theoretical reasoning where form, as such, is pretty much irrelevant. So the explanation of the power of formal reasoning is special; it is that the form of the inference guarantees the truth of the conclusion, if the premises are true.

In my terms, this is still a case in which the considerations adduced favour believing the conclusion; though with formal reasoning we have the special feature that, as we might say, the considerations adduced favour their conclusion conclusively. It is in the explanation of this fact that formal reasoning is special.

Now the form of an instance of formal reasoning can be understood as a relation between propositions. Given p and *if* p *then* q, we can infer that q; the truth of q is guaranteed by the truth of p and *if* p *then* q; this is so whatever propositions we put in the p-place and the q-place, as long as we do it consistently. The explanation of the ability of the premises to favour the conclusion appeals to a relation between propositions. But I maintain (see section 2) that no proposition favours anything at all. It is only matters of fact that can count in favour of one conclusion rather than another. And a true proposition is not the same thing as a matter of fact. A proposition is a representation, and true proposition is still a representation, while a matter of fact is what a true proposition represents. A representation, whether true or false, cannot count in favour of anything. It is only the thing represented that can do that.

But this all fits. Even in modus ponens, it is the things believed, that p and that *if* p *then* q, that together favour believing that q. Someone who reasons from these things does believe them – at least, she does if this is serious reasoning rather than just pretend reasoning in a classroom. If these things believed, these matters of fact, are true, they favour believing that q and do it, one might say, conclusively. Yes, what explains the ability of these two matters of fact to favour believing that q is a relation between propositions, but this does not change the fact that it is those matters of fact that favour believing that q, rather than some relation between propositions. So even in the case of formal reasoning, our basic structure holds good. In reasoning, we move from considerations adduced to the conclusion most favoured by those considerations. But what explains the ability of those considerations to favour our conclusion will

differ from case to case. In practical reasoning, it is some relation to values; in non-formal theoretical reasoning, it is (normally) some relation to probability; and in formal reasoning, it is the formal relations between certain propositions.

9 Intermediate summary

The picture that has emerged is Aristotelian in one central respect, by maintaining that action can be as direct a response to reasoning as can belief. Such differences as there may be between action and belief are admitted but absorbed, supposedly. What is stressed is their similarities. Both are direct responses to reasons given by the considerations adduced, taken together. In reasoning, we determine what form of response is most favoured by those considerations and respond accordingly, whether that response be belief, or action – or doubt, or hope, or any other response that can be favoured by complex combinations of considerations.

Such differences as there are between practical and theoretical reasoning, and between what we might call ordinary theoretical reasoning and the unusual case of formal reasoning, are pretty much all located in answers to the question what explains the ability of the considerations adduced to favour believing the conclusion or acting in the relevant way. Beyond that, we can tell pretty much the same story right across the board. In reasoning, we move from considerations adduced to response, and a practical response can be as directly related to the considerations adduced as can any theoretical response.

10 Getting it wrong

What sorts of mistakes can be made in reasoning? Basically, there are two forms of mistake available. One can reason from things that are not the case, mistakenly taking them to be the case, and one can mistake the relevance of the things one is reasoning from. But I would say that only the latter is properly thought of as a mistake one makes *in* one's reasoning. The former is a mistake too, but the mistake is made before the reasoning begins.

There are of course other sorts of error in the offing. One can fail to collect sufficient relevant information, for instance. But that is not a mistake. It is a defect of a different sort. If you have inadequate information, you may reason well from what you have, just as you can reason well from misinformation. But you will be lucky if your response actually fits the situation confronting you.

I have maintained that in reasoning, we attempt to determine what response is most favoured by the considerations we adduce and to make that response – whether it be a belief or an action. I have also maintained that only things that are the case can favour anything. But of course we can and do reason from things that are not the case; we do this in any case in which we are mistaken about something. Is this an objection to my picture? No, it is not. The person doing the reasoning is reasoning from things that, as she supposes, are so. Perhaps she correctly determines what sort of response would be most favoured by those considerations and responds accordingly. But in fact she is quite wrong about some of the things she is reasoning from. Still, it is for her as if they were so, and she reasons from them in that light. There is no mystery here, no pseudo-favouring being done by things that are not the case. The reasoning is good reasoning because she correctly determines what these things would have favoured, taken together, had they been the case.

I end by considering three difficulties for the story I am trying to tell.

11 First difficulty

The neo-Aristotelian picture is of course vulnerable to complaints that, though belief and action may be similar in certain respects, they are very dissimilar in others. And those dissimilarities might be sufficient to undermine the claim that action can be a direct response to considerations adduced, in the way that belief can.

One such dissimilarity, supposedly, is that one cannot believe at will, in the way that one can act at will. Now first, I believe that one can decide that something is so (which amounts to deciding what to believe), just as one can decide what to do. And second, there are many things that one can do but not do at will – as in any case which requires effort or needs help. But, putting that aside, the important thing is that this difference actually tells in favour of my picture. For I am stressing the active nature of belief, in the analogy with action, while the supposed difference is that belief just happens to one, whether one wants it or not. In this last picture, the idea that in reasoning we are active is a mistake. All that one can do is to marshal the relevant considerations and then hope that the right belief occurs in one. I do not recommend this picture, but it is the consequence of an exaggerated distinction between belief and action.

Still, there do remain differences between belief and action, and these may be sufficient to persuade one that reasoning directly to action is impossible. Joseph Raz stalwartly maintains (most notably in Raz 2011, ch. 6, but also in his 2015) that the most that practical reasoning can do is to serve up a belief about what one has most reason to do, what one ought to do or some other such form of normative belief. In this picture, action may follow reasoning, and follow it immediately, but it cannot be part of that reasoning. One argument is that they may tie you down: you reason well, but just as you are about to act, they tie you down and prevent you from acting. Does this show that your reasoning was incomplete? No: but the fact that you are not at fault – not at fault rationally, that is – supposedly shows that the reasoning itself cannot include the action that you failed to do. You did everything that reason required of you, but you did not act. Your reasoning, therefore, must have been completed even though the relevant action did not get done. So the conclusion of that reasoning must be some such thing as that you ought to act thus, or have most reason to act thus.

My own view about this is that the same may happen with belief. Suppose that you are doing your theoretical reasoning (calculating your probable pension, perhaps), and just when you are about to come to a conclusion, your grandchildren rush in, all ready for the promised trip to the zoo. Are you at rational fault for yielding to the

interruption? I doubt it. But this does nothing to show that the relevant belief was not the proper conclusion of the reasoning.

12 Second difficulty

The second difficulty is how we are to conceive of an action in all this. For there are some conceptions of action that seem to be at odds with any idea that an action could stand in the same relation to reasoning as belief can stand. If an action is a mere motion of the body, all one could say is that such a motion could end reasoning, but it could hardly stand in the more normative relation to reasoning that we are alluding to when we speak of a conclusion. And if an action is a motion of the body caused in a certain way (most plausibly by some combination of beliefs and desires of the agent's), even then it seems hard for it to stand in that normative relation to considerations adduced.

Luckily there are other conceptions of action, more suited to the needs of the theory of reasoning, deliberation and choice. Agent-causation theories come in two styles. The first is that agents cause their own actions. Sometimes this is presented as a theory of free action and not of all action, but not always. Anyway, we should contrast it with a quite different theory which maintains that an action is an agent causing a change (see especially Alvarez and Hyman 1998). Here what is caused is not the action but a change, and the action is the causing of that change. So opening the door is causing the door to become open.

All I will say here (since this is a huge and independent topic) is that I find this last conception of action by far the most plausible, for independent reasons. And in its terms I see no difficulty in supposing that an action can be rationally related to the considerations in light of which it is done. Any action that is done for reasons will count as intentional. Note, however, that this sort of intentionality refers to what is called an 'intention-in-action' rather than to the existence of a 'prior intention'. Those who think that practical reasoning can only lead us to the formation of an intention seem to have a prior intention in mind: I reason to that intention, or plan, and then when the time comes, I act accordingly. I allow, of course, that this does happen. But I also insist that an action to which one reasons will, when one does it, be intentional in a way that does not demand any prior, or intermediate, intention. An intentional action of this other sort cannot be broken up into a physical motion and a prior mental state. Indeed, it is very hard to drive a wedge between the intentionality of the action and the action itself. The sort of watchfulness and control which infuses my action is not a separate or separable mental accompaniment to what is itself a merely physical event. (See McDowell 2015.)

13 Third difficulty

Up to now I have talked blithely about doing the action that is most favoured by the considerations adduced in reasoning. But there is a problem about this, which is why I have sometimes restricted myself to talk about acting *in the way* that is most favoured by those considerations. I associate this problem with H.A. Prichard (2002, ch. 9.viii). The problem is twofold. First, until the action is done, there is no such action as the one most favoured by anything. Second, even if there were an action to be the one most favoured, there are many different ways of doing that action, all of which are equally favoured. It may not matter whether the action is done today or tomorrow, sullenly or willingly, with the right hand or the left and so on. So how can I say that the conclusion of practical reasoning, if things go well, is *the action most favoured* by the considerations adduced? It looks as if all that can be favoured is, not *an* action, but something like a way of acting, or acting in a certain way. The reasons I adduce, taken together, favour my acting in a reimbursing way (since they are reasons to pay back the money) but are silent as to pretty much every detail of when, where, and how.

What then is it to favour 'acting in a way'? I take this to be an unresolved problem in the philosophy of action. Is it the same as favouring *a way of acting*? I am not sure that this expression even makes sense.

The sharp edge of this, as far as my account of practical reasoning is concerned, is that I have spoken repeatedly of considerations as favouring action, and of doing *the action* most favoured by those considerations. And there is no such action.

But this would only be a difficulty for my view in particular if no such difficulty arose for the opposing view that reasoning can only take us to belief, never to action. But what we see is that, though this has not been noticed, the very same difficulty arises for belief. When we speak of certain considerations as reasons to believe that p, and thereby as favouring believing that p, there may as yet be no believing that p to be so favoured. And, what is more, it seems to be the case that there are many different believings that p that are equally favoured by those considerations. For after all, though I may have perfectly good reasons to believe that p, those reasons do not tell me when or where to believe that p, with what enthusiasm to believe it, with what confidence to believe it, and so on. All that the reasons serve up is a partial blueprint for believing: it is to be a believing that p. The rest is left up to me, and perfectly properly.

One might hopefully suppose that believing that p is somehow particularised by its content, the thing believed, which is indeed particular enough. But sadly that does

not succeed in particularising the believing. It remains stubbornly the case that there are various differences between different believings that p, just as there are various differences between different actions of the same general type. And so the Prichard point applies on both sides.

So, even though the Prichard problem is yet to be solved, I maintain that it is not in any way a problem specially for my view of practical reasoning. It applies to any theory of reasoning, practical or theoretical.

14 A final challenge

What follows is not a difficulty for my Aristotelian picture especially but a general challenge for theories of reasoning, which surely earns it a place in a handbook on practical reason. How are we to understand reasoning from hope to hope, or from doubt to doubt? An instance of reasoning from hope to hope might be this:

> I hope that my daughter gets home safely tonight
>
> If she caught the train, she will get home safely tonight
>
> So I hope she managed to catch the train.

One possibility is to deny that this is reasoning. But I don't find that a very plausible escape route. Surely we can do better than that. Perhaps, then, we could turn the whole thing into an inference from belief to belief, in some such way as this:

> It is to be hoped that my daughter gets home safely tonight
>
> If she caught the train, she will get home safely tonight
>
> So it is to be hoped that she managed to catch the train.

This device converts reasoning from hope to hope into reasoning from belief (something believed, that is) to belief. But, though there is nothing wrong with the second passage of reasoning, it seems to me rather strained to say that it is what is really going on in the first.

Putting that aside for a moment, let us think about reasoning from doubt to doubt.

> I doubt that q
>
> If p then q
>
> So I doubt that p

What is causing the problem here is that in reasoning from belief to belief, one can and should keep the 'I believe that' out of the premises. Reasoning from p to q is not reasoning from 'I believe that p' to 'I believe that q'; that would be a different inference. But one cannot extract the 'I doubt that' from the premise and the previous conclusion without total distortion. The person who doubts that p is exactly not going to reason from p to anything else at all.

I confess that I do not know how to make progress with this issue, but would claim that nobody else does either.

References

Alvarez, M., and Hyman, J. (1998) 'Agents and Their Actions,' *Philosophy* 73 (284), 219–45.

Dancy, J. (1981) 'On Moral Properties,' *Mind* xc, 367–85.

Dancy, J. (1983) 'Ethical Particularism and Morally Relevant Properties,' *Mind* xcii, 530–47.

Dancy, J. (1993) *Moral Reasons* (Oxford: Basil Blackwell).

Dancy, J. (2000) *Practical Reality* (Oxford: Clarendon Press).

Dancy, J. (2018) Practical Shape: A Theory of Practical Reasoning *(Oxford: Clarendon Press).*

Davidson, D. (1963) 'Actions, Reasons and Causes,' reprinted in *His Essays on Actions and Events* (Oxford: Clarendon Press, 1980), pp. 3–19.

Kearns, S., and Star, D. (2009) 'Reasons as Evidence,' in R. Shafer-Landau ed. *Oxford Studies in Metaethics* (Oxford: Oxford University Press), Vol. 4, pp. 215–42.

McDowell, J. (1998) *Mind, Value, and Reality* (Cambridge, MA: Harvard University Press).

McDowell, J. (2015) 'Acting as One Intends,' in J. Dancy and C. Sandis eds. (Oxford: Wiley-Blackwell, 2015), pp. 145–58.

Prichard, H. A. (2002) *H. A. Prichard: Moral Writings*, ed. J. MacAdam (Oxford: Oxford University Press).

Raz, J. (1975) 'Reasons for Action, Decisions and Norms,' *Mind* lxxxvi, 481–99.

Raz, J. (2011) *From Normativity to Responsibility* (Oxford: Oxford University Press).

Raz, J. (2015) 'Normativity: The Role of Reasoning,' *Philosophical Issues* 25. Normativity ed. R. Neta.

Thomson, J. J. (2008) *Normativity* (Chicago and La Salle, IL: Open Court) esp. Addendum 4.

15. The guise of the good

Sergio Tenenbaum

According to the "The Guise of the Good Thesis" (henceforth GG), when we act intentionally, we always act under the guise of the good. That is, in φ-ing intentionally, we take φ-ing to be good. Although I will use this formulation, there are many possible variations of the view, and many of its advocates extend the view to apply to any desire; that is, they take it that all desires present their objects as good (Tenenbaum, 2007, Stampe, 1987) or that they are "experiences of value" (Oddie, 2005). In this entry, I outline various versions of the GG view and motivations for it, together with arguments that have been presented for and against this view.

Background and motivation

In very broad terms, advocates of GG will typically defend the view in the course of understanding intentional agency as expressing, or being guided by, our rational or cognitive powers. Although the view is called "Guise of the Good", a similar thesis could be upheld in terms of reasons (Gregory, 2013) or fittingness if, for instance, one were to be committed to the primacy of these normative notions. Versions of GG go back at least to Plato: in the Protagoras, for example, Plato seems to argue against the possibility of some form of *akrasia* on the basis of the impossibility of willingly refraining from pursuing what he knows to be good.[1] Jessica Moss argues that Aristotle was committed to a wide ranging version of GG, namely that "all motivation involves an appearance of the desired object as good" (2010, p. 3). GG has been quite popular in the history of philosophy: indeed, Kant describes a version of GG with respect to desire and motivation as an "old formula of the schools."[2] Kant himself takes this "old formula of the schools" to be "indubitably certain," at least insofar as we will "under the guidance of reason" (1997, p. 52). Uriah Kriegel argues that GG plays a central role in Brentano's theory of intentionality (2018, Part III). Despite their radically different views, G.E.M. Anscombe and Donald Davidson, arguably the two "parents" of contemporary action theory, both seem to have endorsed some version of GG. According to Anscombe, "the question 'What do you want that for?' arises until we reach the *desirability characterization*, about which 'what do you want that for?' does not arise" (2000, p. 74, emphasis mine). Davidson takes intentions, including intentions in action, to be "all-out" evaluative judgments or judgments of desirability (1980b, 1980c). Of course, GG has also had well-known critics: both Hobbes and Hume seem to have rejected the doctrine. According to Hobbes, rather than representing the object of desire as good, "whatsoever is the object of any man's appetite or desire" is what "he for this part calleth good" (1994, Book I, Chapter 6). According to Hume, "A passion is an original existence, or, if you will, modification of existence, and contains not any representative quality" (1978, p. 415). In contemporary philosophy, Michael Stocker and David Velleman have advanced prominent challenges to GG (Stocker, 1979; Velleman, 1992), and more recently, Kieran Setiya has presented novel arguments against the position (2007, 2010).

[1] (Plato, 1991). The dialogue seems also to imply that one pursues only what one regards as good.

[2] Kant describes the "old formula of the schools" as "*Nihil appetimus, nisi sub ratione boni; nihil aversamur, nisi sub ratione mail* (we only desire under the guise of the good; we only avoid under the guise of the bad) (Kant, 1997, p. 51).

Often, GG is defended as preserving some intuitive claims about the nature of desire and intentional action; more specifically, the view seems to capture how explanations of intentional action make the action intelligible or display it in a reasonable light. Anscombe famously argued that intentional action is characterized by a special sense of the question "Why?"; a proper answer to this question provides a desirability characterization; that is, a characterization of the end of the agent that shows that the agent's pursuit is an intelligible. It seems plausible that a pursuit is made intelligible by showing how the agent regarded the object of the pursuit as good, at least in a broad sense of 'good'. If one is asked "Why are you pumping water?" and they answer "So that the water can go a further mile south", it would seem natural to regard this explanation is incomplete and ask "But why do you want the water to go South?" On the other hand, if the answer were "So that the water can reach the children, who'd otherwise have no access to clean water", no further explanation is necessary. It seems that this difference can be accounted for by the fact that the latter explanation, but not the former, shows the good that the agent saw in the action, given that "providing children with clean water" is something that is readily intelligible as the pursuit of a good. GG puts restrictions on what the agent can pursue intentionally; that is, it holds that she can act intentionally only when she conceives her action to be good. In the absence of such restrictions, it seems that in principle anything could be an object of pursuit or desire for any particular agent. It is thus tempting to test the hypothesis by looking into cases in which an object of desire or pursuit does not seem to be the kind of thing that can be conceived as good or in which it is somehow stipulated that the object of pursuit in no way connects to the evaluative judgments of the agent.[3] In Anscombe's famous example, an agent who would answer such a "Why" question by mentioning a desire for a saucer of mud would fail to make her action intelligible. Although it is possible to argue that in such an example the agent is intelligible because the complement of "desire" is a noun, rather than an infinitival or a proposition, the same cannot be said of other examples. For instance, Anscombe considers someone who "hunted out all the green books in his house and spread them out carefully on the roof". If in answer to this version of the question "Why?", the agent simply said "for no particular reason" or "I just thought I would", we would find his answer "unintelligible" (2000, p. 26). Warren Quinn gives the example of someone who has a disposition to turn on radios wherever he encounters them. The agent in question is not interested in listening to music or in anything else connected to having the radio on. Quinn argues that actions of someone moved by such brute dispositions cannot be rationalized at all (and thus arguably it would not count as intentional action at all [1993]).

However, these examples point out to a more general consideration in favour of GG, namely the idea that what is special about intentional action, rather than any other

[3] Of course, one could think that there are different restrictions on the objects of desire and pursuit. But insofar as the following examples are cases in which we think that there is something awry because the objects of pursuit are not, and perhaps could not, be conceived as good, they are effective examples against such views as well.

action that might have originated within a rational agent, is that intentional action is guided by the agent's understanding that the action, or the intended object, was worth pursuing (or, in more modest versions of the view, that there was something in the action or object that was worth pursuing). As Joseph Raz puts it: "From its earliest origins, whatever version of the Guise of the Good was viewed with favour was the keystone keeping in place and bridging the theory of value, the theory of normativity and rationality and the understanding of intentional action" (2010). One way to make this motivation more concrete is to draw an analogy between the role of 'good' in practical rationality and intentional agency and the role of 'true' in theoretical rationality and epistemic attitudes or epistemic agency (Tenenbaum, 2006a, 2008). It is widely accepted that in believing a certain proposition, the subject somehow holds or takes the content of the proposition to be true.[4] Moore's paradox ("it's raining outside, but I don't believe it"), for instance, supposedly illustrates the fact that there is a tension between asserting the truth of a proposition and failing to believe the proposition. Relatedly, truth is taken to be a constitutive standard of correctness for belief, so that an ideal "theoretical agent" only believes what is true (or all her beliefs are instances of knowledge), and an ideally rational believer forms and revises beliefs in accordance with the norms of evidence. In this picture, truth is the formal object of belief (any content that we believe is taken to be true)[5] as well as the formal aim of belief (any belief is correct only if it is true). Of course, each element of this picture of belief can be disputed, but it gives advocates of GG a general model of how to understand the relation between action and the good in terms of the relation between belief and truth (more on this later). In the good case, a subject employs her rational and cognitive faculties to form a belief that p in response to the fact that p; in other words, she believes p because p is the case (is true) and her cognitive faculties provide her with access to the fact that p. The bad case is explained as a failed or impeded exercise of these capacities; an explanation of a belief that falls short of knowledge would explain why the agent takes it to be the case that p even though p is not the case or the fact p is not appropriately connected to this exercise of the subject's cognitive faculties.

A parallel view about GG argues that in intending to A or acting with the intention of A-ing, an agent takes A-ing to be good, or believes that A-ing is good. Similarly, in some versions of GG, GG takes the formal object and the formal aim of intentional agency to be the good; in A-ing intentionally or intending to A, an agent takes A to be good and the internal standard of success in action (and intention) is that the agent does (and intends) what is in fact good. In the good case, intentional agency is explained simply as the exercise of one's (practical) cognitive and rational faculties employed in the pursuit of the good. And the bad case will be similarly a case in which

[4] The first sentence of Eric Schwitzgebel's Stanford Encyclopedia entry on "belief" is: "Contemporary analytic philosophers of mind generally use the term "belief" to refer to the attitude we have, roughly, whenever we take something to be the case or regard it as true" (Schwitzgebel, 2019).

[5] Or taken to be a representation of the world as it is.

the exercise of these faculties was somehow defective. This version of GG promises also to vindicate the way in which intentional explanations show an agent's actions to be "intelligible" (Anscombe, 2000) or the way in which intentional explanations show how "from the agent's point of view there was, when he acted, something to be said for the action." (Davidson, 1980a, p. 9). But more generally, according to GG, explaining why the agent performed a certain action amounts to explaining the good that she saw in the action she undertook (or in its consequences).

Although this presentation of GG relies on the analogy with the role of truth in understanding belief and knowledge, the two pictures are independent. One could accept GG without committing oneself to the parallel picture about the nature of belief or knowledge. However, rejecting the equivalent picture in the theoretical realm would rob GG of an important motivation, namely its ability to provide a unified understanding of practical and theoretical cognition and rationality in which theoretical and practical rationality are distinguished only by having different formal objects or aims.

It is worth mentioning some related motivations or arguments for GG. A more neo-Aristotelian approach to GG sees this thesis as essential for understanding how a general notion of the good that is constitutive of the teleological nature of all life forms can be extended to the realm of rational agency. In this picture, animals act in pursuit of ends that are naturally good for them in light of their form. A cat hunts mice because it is of the nature of cats to hunt mice; hunting mice is good for a cat insofar as in hunting mice the cat actualizes its form. When a cat hunts a mouse, it acts successfully in pursuing a goal insofar as it has this goal in virtue of its form. Human agents are moved by a *self-conscious* representation of their good. If a capacity for action in the animate world in general is the capacity to pursue the animal's good, the rational, self-conscious, capacity of a rational agent to act is "a capacity to pursue what it takes to be good" (Boyle & Lavin, 2010, p. 187).

Another possible motivation flows from a commitment to value realism. In a certain interpretation, GG takes our conative faculties to provide access to an independent realm of value; in this picture, desires are experience of value (Oddie, 2005). Of course, there need not be a single, or even a very general, main reason to accept GG. Benjamin Wald, for instance, argues that we should accept GG because of its overall fruitfulness in various branches of practical philosophy (Wald, 2017).

The nature of the guise

There are various versions of GG, but here I will focus on a very basic distinction between two versions of the "guise". In one version of GG, an agent acts intentionally only if she *believes* that her action is good. So, for instance, Raz presents the following as an immediate consequence of the defining theses of GG:

> Intentional actions are actions taken in, and because of, a belief that there is some good in them.
>
> (2010, p. 111)

In this view, roughly, our reasoning about what to do is reasoning that ends in the belief that a certain course of action is good and intentional action is (typically)[1] action done in light of, and because of, this belief. We can call this version of GG the "Content Version", in light of the fact that the "guise" appears as the content of an attitude, namely the belief that explains the action. In an alternative view, in having an intention (either a future-directed intention or an intention an action), an agent takes the object of the intention to be good. If we extend this version of GG to desires, this view says that in desiring, the content of the desire appears to be good to the agent, while extending the Content view to desire yields a view in which desire is an appearance with an evaluative content or a belief that something is *pro tanto* or *prima facie* good. We can call this version the "Attitude version", given that the "guise" does not appear in the content of any of the agent's attitudes, but it is part of the nature of the attitude itself that in having such an attitude we somehow take, or hold, its content to be good.[2]

The most obvious advantage of the Content version is that it does not need to rely on the notion of "taking" or "holding" that is essential for the Attitude version. After all, what is taking X to be good if not a belief that X is good? However, even the Content version is committed to a similar "taking" relation if we accept the view mentioned before that to believe that p is to take p to be true. The "taking" relation is already needed in order to understand the relation between belief and the truth; the Attitude version simply postulates that the same relation holds between intending to A, or A-ing intentionally, taking A to be good. In fact, in the Attitude version, there is a much

[1] Raz allows that some actions are not done under the guise of the good. See (Raz, 2010, 2016).

[2] For versions of this view, see (Kriegel, 2018; Schafer, 2013; Tenenbaum, 2018, 2008, 2012; Wald, 2017). (Velleman, 1992) makes a similar distinction, but the paper ultimately rejects GG.

clearer parallel between the realms of theoretical and practical rationality: the role of 'good' in practical reasoning and action is the same as the role of 'true' in theoretical reasoning and belief. How exactly the parallel is spelled out will depend on the specific theory;[3] here, I can only outline some of these parallels. For instance, good and true may play similar roles in distinguishing various kinds of 'practical' and 'theoretical' attitudes. We have, on the one hand, unendorsed, or *prima-facie* attitudes towards the good and the true, such as, respectively, desire and perceptual appearances, and on the other hand, endorsed, or *all-out* ones, such as intentions, or intentional actions, and belief. The different formal objects may also play similar roles determining what counts as valid inference (good-preserving in the case of practical reasoning; truth-preserving in the case of theoretical reasoning). Finally, they may also play similar roles in determining the fundamental case of success in belief and action; in other words, what counts as theoretical and practical knowledge (non-accidental true belief in the case of theoretical reasoning, non-accidental good action in the case of practical reasoning).

The Content version also seems to require richer conceptual capacities from those who can engage in intentional action. Only those who have evaluative beliefs can act intentionally, and for every intention and intentional act, there must be a corresponding (albeit implicit) belief with the relevant content. This is particularly problematic if we want to extend GG to desire, since it would imply that small children and animals do not have desires, or at least not the same kinds of desires we have. Of course, one might restrict GG to intention and intentional action, but not extending GG to desires leaves GG incapable of explaining the 'rational force' of desires;[4] that is, it puts GG in a difficult position for accounting for the role of desires in practical reasoning. A more promising route would be to argue that the evaluative content of the relevant desires is non-conceptual and thus capable of figuring in the content of children's and animals' mental states.[5]

Similarly, Content versions of GG will arguably face more difficulties in explaining apparent cases in which the agent seems to act contrary to their evaluative beliefs, such as cases of *akrasia* and perversion (more on these issues subsequently). Finally, in the Content version, practical reasoning turns out to be an instance of theoretical reasoning, albeit theoretical reasoning whose conclusion is about the good.[6] On the other hand, the Attitude version, practical reasoning is a genuinely different form of reasoning: it is a form of reasoning, whose soundness or validity cannot be understood in terms of truth-conduciveness or truth-preservation. Of course, this difference does

[3] For one example, see (Tenenbaum, 2007).

[4] See (Schafer, 2013).

[5] See (Hawkins, 2008).

[6] Anscombe argues that in such a view, there is nothing that there should be called a 'practical syllogism': just as we see no reason to think that 'mince-pie syllogisms', syllogisms whose subject-matter are mince-pies, express a different form of reasoning, we have no reason to think, in this view, that the practical syllogism is a special form of syllogism. (Anscombe, 2000, p. 58)

not seem to favour either view on its own. However, philosophers have recently argued that a proper understanding of intentional action requires that reasoning reach all the way to the actual actions of an agent; the action itself must be the conclusion of practical reasoning or the direct expression of our rational powers.[7] If this is correct, this would be a further reason to accept the Attitude version of GG, given that, in the Content version, practical reasoning ends at the formation of the relevant belief.

[7] On this point, see (Lavin, 2013; Tenenbaum, 2006b).

Common objections and replies

In the last few decades, GG has been the subject of a large number of criticisms and objections. In his seminal paper in the topic, Michael Stocker says:

> It is hardly unfair, if unfair at all to suggest that the philosophical view is overwhelmingly that the good and only the good attracts.
>
> (1979, pp. 739–740)

Stocker would have been pleased to learn that this state of the discipline that he so clearly lamented has been radically changed: we now have no shortage of philosophers who either explicitly reject GG or who provide accounts of human agency that are incompatible with GG. Opponents of GG often argue against the view by proposing putative counterexamples to it (Stocker, 1979; Velleman, 1992). We'll briefly examine three central types of purported counterexamples to GG, namely cases of perversion, cases of *akrasia*, and cases of "arational" actions.

Purported counterexamples: Akrasia

Perhaps the most common counterexample raised against GG are cases of weakness of will or *akrasia*.[1] To make this challenge clear, it is worth taking a step back and looking at a very basic objection to GG. According to this objection, GG must get the structure of motivation wrong, since we may want or be motivated to do things that we don't believe to be good in any way. So, for instance, Gary Watson gives the example of "a squash player who, while suffering an ignominious defeat, desires to smash his opponent in the face with the racquet" (Watson, 1975, p. 210). An agent might have this motivation and yet not find it to be good in any way to behave in this manner. A common response to this objection is to say that in such cases, the want corresponds to a *prima-facie* evaluative judgment, or a perception or appearance of value;[2] this is certainly compatible with the agent believing that the object of the desire has no value. We can compare such cases with perceptual illusions in the theoretical realm. Sticks

[1] Richard Holton takes weakness of will and *akrasia* to be different phenomena. What Holton describes as weakness of will does not present a particular problem for GG; the issues that we discuss subsequently fall under the heading of what Holton calls "*akrasia*" (Holton, 1999).

[2] For understanding desires in terms of *prima-facie* evaluations, or appearances or experiences of the good, see (Davidson, 1980b, 1980c; Oddie, 2005; Stampe, 1987; Tenenbaum, 2007).

might look bent under water, cars might look small from a distance, and the lines in the Müller-Lyer illusion appear to be of different sizes.[3] These things continue to *appear* this way, even when we know that the stick is straight, that the cars are large, or that the lines are of the same size. Similarly, we can continue to desire to smash the racket on our opponent (smashing the opponent with the racket *appears* good), even when we know there is no value in doing so. In other words, the relation between desire and intention, or intentional action, is like the relation between perceptual (and other) appearances and belief: desire is an unendorsed, or *prima-facie*, attitude towards the good whose existence is compatible with the absence of any instance of an endorsed, or *all-out*, attitude with the same content.

But this move does not necessarily respond to the challenge presented by cases of *akrasia*. Weak-willed agents do not simply desire that which they regard as worthless or less valuable than an alternative. Weak-willed agents *pursue*, or at least form intentions to pursue, actions that they consider worse than alternatives open to them. Their "endorsed", *all-out* attitudes seem to favour the action that they regard to be bad (or at the least worse). But if the agent acts intentionally under the guise of the good, wouldn't she always prefer a better option over a worse option? The weak-willed agent seems to choose, say, to watch a full season of her favourite series instead of studying for her exam, while simultaneously believing that watching the series is the worse option. Some philosophers sympathetic to GG think that all that GG requires is that an agent act in the pursuit of *some good* (but not necessarily the better option),[4] but akratic agents might engage in actions that they do not think are good in any way; Watson's squash player might succumb to temptation and attack his opponent. Davidson himself took *akrasia* to be a serious challenge to the claim that intentions in action are evaluative judgments. In a seminal paper (Davidson, 1980b),[5] Davidson distinguishes between "all-things-considered" evaluative judgments and "all-out" evaluative judgments. An all-things-considered judgment is an evaluative judgment that takes into account all the relevant considerations. So in our example, the akratic agent forms the judgment "given all the relevant considerations, it is best to study". However, this is not an unconditional judgment of what it is best to do *simpliciter*, so it is compatible with the other judgment that on Davidson's view the akratic agent makes, namely an "all-out" unconditioned judgment. In our example, the agent also makes the following judgment: "it is best to watch TV". Of course, our unconditional evaluative judgments should take into account all the relevant considerations, so the agent in question is guilty of irrationality. This is, however, a welcome consequence: *akrasia* is a form of irrationality.

[3] (Austin, 1962) has famously argued that the stick under water does not really look bent. I'll ignore such complications, and, at any rate, a similar claim would be rather implausible regarding the Müller-Lyer illusion.

[4] See (Clark, 2010). For a criticism of this more moderate version of GG, see (Tenenbaum, 2009).

[5] For further development of this account of *akrasia*, see (Tenenbaum, 1999).

Some philosophers find this move unpersuasive; they argue that the akratic agent often acts against their "all-out" judgment of what is best (Bratman, 1979; Pears, 1982; Tappolet, 2003) or that there is no non-perspectival comparative judgment that the agent makes in favour of the akratic action (McDowell, 2010). But it is not clear that the advocate of GG need to accept that the agent makes an all-out comparative judgment. One can insist that the relevant judgment is a judgment about what is *good simpliciter* and thus argue that the akratic agent moves from a defeated or undermined appearance or *prima-facie* judgment to the conclusion that something is good *simpliciter*, much in the same way as a subject could irrationally move from defeated or undermined evidence to an irrational belief (Tenenbaum, 2018). At least if one accepts the Attitude version of GG, this stance is compatible with the agent still *believing* that it is best to study. After all, the evaluative judgment in question, the way in which the weak-willed agent regards watching the show to be good, is not a belief.

Purported counterexamples: perversion

Cases of perverse action (Stocker, 1979, 2008; Sussman, 2009; Velleman, 1992) and cases that fall under Hursthouse's category of "arational actions" (1991) also seem to present difficulties for GG. Perverse actions are actions that are done exactly because they are bad, rather than being good. Satan is supposed to be an illustration of an agent who performs actions because they are bad. Anscombe, however, argues that we can make sense of the good pursued by Satan:

> the good of its being bad ... might be condemnation of good as impotent, slavish, and inglorious. Then the good of making evil my good is my intact liberty in the unsubmissiveness of my will.
>
> (Anscombe, 2000, p. 75)

Moreover, it is not clear that Satan pursues what is bad *simpliciter*, rather that what is *morally* bad or some other specific form of badness. After all, Satan does not seem to find anything attractive in foul-tasting food, badly played music, or being engaged in boring activities, even though all these things are also bad. If a perverse agent is attracted by *badness as such*, why wouldn't she be attracted (at least to some extent) to all instances of badness?[6]

Arational actions are supposed to express an emotion – a jealous lover might, for instance, smash the picture of his beloved, but the lover might not see anything good in a broken picture of the beloved. However, even though the agent need not see the outcome of the action as good (the broken picture), it is not clear why one needs to deny that he might see the action itself (the breaking of the glass) as good (Boyle

[6] Tenenbaum (2018). For other responses to this objection, see (Raz, 2016, 2018).

& Lavin, 2010; Tenenbaum, 2007).[7] In sum, purported counterexamples might lead the GG advocate to refine and qualify the view, but, given its central theoretical motivations, it is unlikely to present insurmountable problems for the view.[8]

Theoretical difficulties: alternative constitutive aims

Most of the authors who raise such counterexamples also try to argue more systematically that a proper understanding of intentional agency does not require GG. These strategies are mostly of two kinds: either they dispense altogether with the idea that there is a formal aim or object that is constitutive of intentional agency (Setiya, 2007), or they propose a different a constitutive aim, such a self-understanding or intelligibility (Velleman, 1996).[9] Whether proposals of the latter kind succeed depends on whether the constitutive aim of action is a genuine alternative, and superior, to the one provided by GG.[10]

Although I can't examine in detail here alternative proposals for constitutive aims of action, it is worth mentioning one advantage that GG has over other proposals. The constitutive aim of action needs to do double duty. First, a constitutive aim of action is supposed to be an aim that one necessarily pursues whenever one pursues any other end. So if Velleman is right, whenever I, say, go to mall to buy shoes, I am also pursuing the end of self-understanding or intelligibility. But the constitutive aim is also supposed to provide a normative standard for the action: when I fail to realize the constitutive aim, my action falls short in some important way. But, assuming I can act intentionally and yet fail to realize the end of self-understanding to some degree, why shouldn't I perform an action that provides me with less self-understanding but more of some other end of mine (such as, for instance, personal enrichment). Why shouldn't I sacrifice a bit of self-understanding for a lot of money?

[7] But some advocates of GG think that there are exceptions to the view. The exceptions are typically cases in which the agent has less than full control over her action. See (Raz, 1999, pp. 36–44, 2010)

[8] This point is also made by opponents of GG. See (Setiya, 2010, p. 83)

[9] For further developments of this idea, see the articles collected in (Velleman, 2006) and (Velleman, 2009).

[10] For criticisms of Velleman's proposal for a constitutive aim of action, see (Katsafanas, 2013, Chapter 3). Katsafanas proposes different constitutive aims for intentional action, namely will to power and agential activity. If the "schmagency" objection to Velleman's view works, it would also speak against the idea that there is an interesting notion of intentional action such that it is true that acting intentionally requires us to have a substantive aim such as self-knowledge. See (Enoch, 2006). For a response to the schmagency objection, see (Ferrero, 2009).

Theoretical difficulties: superfluity

Setiya (2010)[11] argues that GG imposes a superfluous constraint on the nature of intentional action. Let us assume for a moment Anscombe's view that an intentional action is one in which I know not only that I am *A*-ing but also why I am *A*-ing – that is, I know my reason for *A*-ing. But the relevant reason here is an explanatory reason: I must know the reason that *explains* my action, not the reason that justifies my action. But even the GG advocate needs to accept that sometimes agents act in ways that are not in fact good. After all, the claim is that agents must *represent* their action as good or believe that their action is good: everyone knows that agents act for bad reasons and thereby pursue actions that are not in fact good. But since being an explanatory reason for an action does not require that the reason be a good reason, neither should knowledge of the reason require that the agent believe that the reason is a good one. Thus, it seems perfectly possible that an agent could know that she is *A*-ing and that she knows that her reason for *A*-ing is that *A*-ing will bring about outcome O and knows that "bringing about outcome O" is a bad reason and thus that O is in no way good. Such an agent fulfils all the conditions of intentional agency even though she's not acting under the guise of the good.

But this argument seems to move from the third-person perspective to the first-person perspective in a possibly illicit way. The fact that someone can rightly explain my action by referring to a reason she knows to be bad does not mean that I can decide on a reason that I know to be a bad reason to act or pursue an action that I do not regard as good. A comparison with Moore's paradox is relevant here: although it is coherent for someone to ascribe a false judgment to me, or to see that I make a judgment based on poor evidence, it is far from clear that I can, at least under normal circumstances, judge that p when I regard p to be false, or judge that p on grounds that I myself take to be inadequate.

[11] (Setiya, 2007) provides a different argument against GG, namely that GG cannot explain certain necessary truths about intentional action. I discuss this argument in (Tenenbaum, 2012).

Conclusion

Critics of GG focused first mostly on purported counterexamples. But the previous discussion hopefully shows that more sophisticated versions of GG can accommodate the phenomena that are supposed to create difficulties for the view. I hope to have also shown that it is far from clear that alternatives can replicate, or dispense with, the theoretical advantages of GG. Although GG has faced a number of criticisms recently, it remains a compelling view of desire and intentional action, a view that can be an important part of a unified account of our rational powers of action and knowledge.

References

Anscombe, G. E. M. (2000). *Intention* (2nd ed.). Cambridge, MA: Harvard University Press.

Austin, J. L. (1962). *Sense and Sensibilia* (G. J. Warnock, Ed.). Oxford: Oxford University Press.

Boyle, M., & Lavin, D. (2010). Goodness and Desire. In S. Tenenbaum (Ed.), *Desire, Practical Reason, and the Good* (pp. 158–199). New York: Oxford University Press.

Bratman, M. (1979). Practical Reasoning and Weakness of the Will. *Nous,* 13 153–171.

Clark, P. (2010). Aspects, Guises, Species, and Knowing Something to Be Good. In S. Tenenbaum (Ed.), *Desire, Practical Reason, and the Good* (pp. 234–244). Oxford: Oxford University Press.

Davidson, D. (1980a). Actions, Reasons, and Causes. In *Essays on Actions and Events* (pp. 3–20). New York: Oxford University Press.

Davidson, D. (1980b). How Is Weakness of the Will Possible? In D. Davidson (Ed.), *Essays on Actions and Events* (pp. 21–43). New York: Oxford University Press.

Davidson, D. (1980c). Intending. In *Essays on Actions and Events* (pp. 83–102). New York: Oxford University Press.

Enoch, D. (2006). Agency, Shmagency: Why Normativity Won't Come from What Is Constitutive of Action. *The Philosophical Review*, *115*(2), 169–198.

Ferrero, L. (2009). Constitutivism and the Inescapability of Agency. *Oxford Studies in Metaethics*, *4*, 303–333.

Gregory, A. (2013). The Guise of Reasons. *American Philosophical Quarterly*, *50*(1), 63–72.

Hawkins, J. (2008). Desiring the Bad Under the Guise of the Good. *The Philosophical Quarterly*, *58*(231), 244–264.

Hobbes, T. (1994). *Leviathan* (E. Curley, Ed.). Indianapolis: Hackett.

Holton, R. (1999). Intention and Weakness of Will. *The Journal of Philosophy*, 241–262.

Hume, D. (1978). *A Treatise of Human Nature, 1739 of 1888* (L. A. Selby Bigge, Ed., revised by P. H. Nidditch, 2nd ed.). Oxford: Clarendon.

Hursthouse, R. (1991). Arational Actions. *The Journal of Philosophy*, *88*(2), 57–68. http://doi.org/10.2307/2026906.

Kant, I. (1997). *Critique of Practical Philosophy.* (M. Gregor, Ed.). New York: Cambridge University Press.

Katsafanas, P. (2013). Agency and the Foundation of Ethics: Nietzschean Constitutivism. *Oxford: Oxford University Press.*

Kriegel, U. (2018). *Brentano's Philosophical System: Mind, Being, Value.* Oxford: Oxford University Press.

Lavin, D. (2013). Must There Be Basic Action? *Nous*, *47*(2), 273–301.

McDowell, J. (2010). What is the Content of an Intention in Action? *Ratio (New Series)*, *XXIII*, 61–78.

Moss, J. (2010). Aristotle's Non-Trivial, Non-Insane View that Everyone Always Desires Things under the Guise of the Good. In S. Tenenbaum (Ed.), *Desire, Practical Reason, and the Good* (pp. 65–81). Oxford: Oxford University Press.

Oddie, G. (2005). *Value, Reality, and Desire.* New York: Oxford University Press.

Pears, D. (1982). How Easy Is Akrasia? *Philosophia*, *11*, 33–50.

Plato. (1991). *Protagoras* (C. C. Taylor, Ed.). Oxford: Oxford University Press.

Quinn, W. (1993). Putting Rationality in Its Place. In W. Quinn (Ed.), *Morality and Action* (pp. 228–255). New York: Cambridge University Press.

Raz, J. (1999). *Engaging Reason: On the Theory of Value and Action.* New York: Oxford University Press.

Raz, J. (2010). The Guise of the Good. In S. Tenenbaum (Ed.), *Desire, Practical Reason, and the Good* (pp. 111–137). New York: Oxford University Press.

Raz, J. (2016). The Guise of the Bad. *Journal of Ethics and Social Philosophy*, *10*(3), 1–15.

Schafer, K. (2013). Perception and the Rational Force of Desire. *Journal of Philosophy*, *110*(5), 258–281.

Schwitzgebel, E. (2019). Belief. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy.* https://plato.stanford.edu/archives/fall2019/entries/belief/.

Setiya, K. (2007). *Reasons Without Rationalism.* Princeton: Princeton University Press.

Setiya, K. (2010). Sympathy for the Devil. In S. Tenenbaum (Ed.), *Desire, Practical Reason, and the Good* (pp. 82–111). New York: Oxford University Press.

Stampe, D. (1987). The Authority of Desire. *The Philosophical Review*, *96*(3), 335–381.

Stocker, M. (1979). Desiring the Bad: An Essay in Moral Psychology. *The Journal of Philosophy*, *76*(12), 738–753.

Stocker, M. (2008). On the Intelligibility of Bad Acts. In D. Chan (Ed.), *Moral Psychology Today* (pp. 123–140). Dordrecht: Springer.

Sussman, D. (2009). For Badness' Sake. *Journal of Philosophy*, *106*(11), 613–628.

Tappolet, C. (2003). Emotions and the Intelligibility of Akratic Action. In S. Stroud & C. Tappolet (Eds.), *Weakness of Will and Practical Irrationality.* New York: Oxford University Press.

Tenenbaum, S. (2018). The Guise of the Guise of the Bad. *Ethical Theory and Moral Practice*, *21*, 5–20.

Tenenbaum, S. (1999). The Judgment of a Weak Will. *Philosophical and Phenomenological Research*, *59*(4), 875–911.

Tenenbaum, S. (2006a). Direction of Fit and Motivational Cognitivism. *Oxford Studies in Metaethics*, *1*, 235–264.
Tenenbaum, S. (2006b). The Conclusion of Practical Reason. In Sergio Tenenbaum (Ed.), *Moral Psychology* (pp. 323–343). Amsterdam: Brill Rodopi.
Tenenbaum, S. (2007). Appearances of the Good: An Essay on the Nature of Practical Reason*. Cambridge: Cambridge University Press.*
Tenenbaum, S. (2008). Appearing Good. *Social Theory and Practice*, *34*(1), 131–138.
Tenenbaum, S. (2009). In Defense of "Appearances". *Dialogue*, *48*, 411–421.
Tenenbaum, S. (2012). Knowing the Good and Knowing What One Is Doing. *Canadian Journal of Philosophy*, 91–117.
Velleman, D. (1992). The Guise of the Good. *Nous*, *26*(1), 3–26.
Velleman, D. (1996). The Possibility of Practical Reason. *Ethics*, *106*(4), 694–726.
Velleman, D. (2006). *Self to Self*. New York: Cambridge University Press.
Velleman, D. (2009). *How We Get Along*. New York: Cambridge University Press.
Wald, B. (2017). *Judging the Guise of the Good By Its Fruits*. Toronto: University of Toronto. PhD Dissertation.

16. Motivational internalism and externalism

Connie S. Rosati

Suppose that Ann judges that she morally ought to keep a promise to her friend Juan to drive him to the airport. Ordinarily, she will be motivated, at least to some degree, to act on her judgment and fulfill her promise. But now suppose that another friend, Malcolm, becomes seriously ill and Ann judges that she morally ought to drive him to the emergency ward rather than drive Juan to the airport. Ordinarily, when a person changes her judgment about what she morally ought to do, she becomes motivated, at least to some degree, to act on her new judgment rather than on her initial judgment. Of course, a competing motivation may, in the end, win out; Ann's desire to see a movie may override her motivation to keep her promise to Juan or even her motivation to drive Malcolm to the emergency ward. Nevertheless, we expect a person who judges that she morally ought to do something to be at least somewhat motivated to do it. A person who judged that she morally ought to do something but who evinced no motivation at all to do it would tend to strike us as either insincere or confused, as not understanding *what it is* to judge that one morally ought to do something.

Philosophers have attempted to explain the strong connection between moral judgment and motivation in two competing ways.[1] According to the view called *motivational internalism*, sometimes also called *motivational judgment internalism*, there is a *necessary* connection between moral judgment and motivation. Necessarily, a person who judges that she morally ought to do something is motivated (at least to some degree) to do it.

Motivational judgment internalism, as the name suggests, is a thesis about the character of moral judgments. As such, it should be distinguished from the view called *existence internalism*. Whereas judgment internalism maintains that there is a necessary connection between moral judgment and motivation, existence internalism maintains that there is a necessary connection between something having a certain normative status and motivation (Darwall 1983: 54–55). Examples of the latter thesis would be the view that a consideration or fact couldn't be a *reason* for a person to act unless it were capable of motivating her and the view that something couldn't be *good for* a person unless she could be motivated by it.[2]

Motivational internalism, as a thesis about the character of moral judgment, contrasts with the view known as *motivational externalism*.[3] According to motivational

[1] I focus in this chapter on motivational internalism and externalism as theses about moral judgment, as this is also the main focus in the philosophical literature, but the debate between motivational internalists and externalists is part of a more general debate about normative judgment, where normative judgments include judgments about goodness, reasons, blameworthiness, and ought judgments.

[2] For defense of existence internalism about reasons, see, for example, Williams (1981) and Darwall (1983). See also, for example, Schroeder (2007: 165–167), but see Parfit (2011, Vol. I: Part 1, Ch. 3). For a discussion of existence internalism about what is good for a person, see Rosati (1996), but see Sarch (2010).

[3] Frankena (1976) and Darwall (1983) attribute the labels to Falk (1952).

externalism, moral judgments motivate *contingently* (albeit rather reliably) because of the presence of some desire that is independent of those judgments, such as the desire to do the morally right thing. The person who judges that she morally ought to φ is motivated to some degree to φ, insofar as she has a desire, for example, to do what she ought. Because the connection between moral judgment and motivation is contingent, the person who judges that she morally ought to φ but fails to be motivated to φ need be neither insincere nor confused about what it is to judge that she morally ought to φ.

This chapter will focus on motivational internalism and motivational externalism – hereinafter, "internalism" and "externalism" – and their competing claims about the connection between moral judgment and motivation.[4] Internalism has been thought by many to bear on determining the correct metaethical theory. Most notably, it has been used in defense of noncognitivist or expressivist theories, according to which moral judgments express conative states (such as approval) rather than beliefs. Some philosophers, though, have defended forms of cognitivism that they take to be compatible with internalism.[5] We shall leave to one side, however, debates about the metaethical implications of the internalism/externalism debate and focus on the latter debate itself, which is of interest in its own right.

After considering a simple formulation of internalism and common externalist challenges to it, we will consider more qualified versions of internalism. Internalism, in its various forms, is ordinarily taken to be an a priori or conceptual thesis about the connection between moral judgment and motivation, rather than an empirical thesis. But as we shall see, some have recently treated it as an empirical thesis, and some have thought that empirical evidence may bear on its truth even as a conceptual thesis.

[4] The single best collection of recent work on motivational internalism of which I am aware is Björnsson et al. (2015). See the introduction for an excellent overview of the internalism/externalism debate and an extensive bibliography. For additional references, see Rosati (2016).

[5] See, for example, Smith (1989).

Motivational internalism

Björnsson et al. (2015: 1) have offered the following characterization of what they call "Simple Internalism":

> Necessarily, if a person judges that she morally ought to φ, then she is (at least somewhat) motivated to φ.[1]

This characterization of internalism is *unconditional* in that it places no restrictions on the circumstances under which a person's moral judgments are motivating; it thus entails that every moral judgment, just as such, is motivating. So let us refer to it as "Unconditional Internalism." Notice that unconditional internalism, and indeed, any version of internalism, requires only that moral judgments motivate *somewhat*. No form of internalism according to which moral judgments provide sufficient motivation to φ would be plausible because of the possibility of *akrasia* or weakness of will.

Externalists have argued that Unconditional Internalism is implausible. They commonly offer, as an alleged counterexample, the conceivability of the amoralist – a person who makes moral judgments while remaining utterly unmoved by them.[2] Because we can conceive of such a person, they contend, it is not true that necessarily, if a person judges that she morally ought to φ, then she is (at least somewhat) motivated to φ.

Internalists have offered various responses to the "amoralist challenge." Some deny that the amoralist is conceivable. Some contend that the amoralist is not really making moral judgments; he or she uses moral terms only in an "inverted commas" sense.[3] Others argue that if the amoralist is making genuine moral judgments, then he must be motivated to *some* degree, it's just that whatever motivation he feels is faint and easily overridden by other motives.[4]

Externalists, of course, find these replies unpersuasive. They ask us to consider, beyond the extreme case of the amoralist, a variety of other cases in which, they say, persons make moral judgments without being at all motivated. Consider, for example, a person who is depressed or apathetic or exhausted or emotionally disturbed and so

[1] I here follow Björnsson et al. in using "to judge," "judgment," and "make a judgment," to "refer to a mental state or mental act" (2).

[2] See, for example, Brink (1989, 1997). For other defenses of externalism, see, Svavarsdóttir (1999, 2006), Shafer-Landau (1998, 2000, 2003), and Zangwill (2015).

[3] See, for example, Hare (1963).

[4] We will examine this last reply in more detail later.

unmotivated by his moral judgments. Or consider a person who may at one point have been motivated by a moral judgment, say, that she morally ought to help the poor, but who, after retiring from twenty years of work on poverty relief, continues to make that moral judgment sincerely without any longer being motivated by it. These cases are all conceivable, externalists contend, and so internalism must be false. Although the latter examples do not themselves involve amoralism, let us use the label "amoralist challenge" to include these varied cases of alleged motivational failure.

Conditional vs unconditional internalism

Internalists have responded to the amoralist challenge by developing forms of *conditional internalism*.

> Conditional Internalism: Necessarily, if a person judges that she morally ought to φ, then she is (at least somewhat) motivated to φ *if she is C*.
>
> (Björnsson et al. 2015: 7)

Internalists have made various proposals as to how internalism should be qualified.

Some have suggested that the relevant conditions are ones in which the person is "psychologically normal" (see, e.g., Björnsson 2002; Timmons 1999). This would exclude cases in which the person making the moral judgments is depressed, emotionally disturbed, exhausted or apathetic. Some have suggested instead that the relevant conditions are ones in which the person making the moral judgment is "morally perceptive" (see, e.g., McDowell 1978, 1979). If a person is morally perceptive, she will both see what she morally ought to do and will be motivated to do it.

Perhaps the most well known proposal is Michael Smith's. According to Smith, we should understand motivational internalism as stating a defeasible connection between moral judgment and motivation, so as to allow, for example, for the possibility of weakness of will:

> If an agent judges that it is right for her to φ in circumstances C, then either she is motivated to φ in C or she is practically irrational.
>
> (1994: 61)

According to this version of internalism, an agent who judges it right to φ is motivated to φ, in the absence of forms of practical unreason or distorting influences on the will.

Smith goes on to offer an internalist challenge to the externalist. According to Smith, the externalist is committed to a problematic picture of moral motivation, which we can see by comparing it to the picture the internalist offers. The internalist explains the connection between moral judgment and motivation as due to the content of moral judgments. The internalist holds that a person is motivated to do the very thing she judges that she morally ought to do, "where this is read *de re* and not *de dicto*" (Smith

1994: 73). For example, the person who judges that she morally ought to aid a stranger in distress acquires and is moved by a nonderivative desire to assist the stranger.

In contrast, the externalist must explain the connection between moral judgment and motivation not by appealing to the content of moral judgments, but by appealing to the "content of the motivational dispositions possessed by the good and strong-willed person" (Smith 1994: 71). The externalist must hold that the person who judges that she morally ought to do something is moved as a result of the motivational dispositions that she has in being a good person.

What might these dispositions be? According to Smith, if the externalist is to account for how our motivations shift with changes in our moral judgments, then the only disposition that could account for such shifts is the motivation to do the right thing, whatever that turns out to be. The good person, the externalist must think, is motivated to do the right thing, "where this is read de dicto and not de re" (Smith 1994: 75). For example, the husband who is forced to choose between saving his wife and a stranger is not motivated by the non-derivative desire to save his wife; rather, he is motivated to save his wife because he is motivated to do the right thing, and because he judges that, in the circumstances, saving her just happens to be the right thing. According to Smith, this picture of moral motivation is implausible and involves a form of "fetishism." The good person, he claims, cares non-derivatively, for example, about another's welfare or about another's being treated justly. To be motivated non-derivatively by concern to do what one believes right and not by non-derivative concern for another's welfare or just treatment is "a fetish or moral vice" (Smith 1994: 75). The externalist, in taking the good person to be motivated to do whatever she happens to believe to be morally right, "alienates her from the ends at which morality properly aims" (76).[1]

Externalists have offered various responses to Smith's charge of fetishism. They have argued that there is nothing fetishistic about a person's being motivationally disposed to do the right thing (Copp 1997: 49–50; Lillehammer 1997: 191–192). In fact, a *de dicto* concern for what is right plausibly plays a critical role in the psychology of the good person (Lillehammer 1997: 192). For the husband confronted with saving either his drowning wife or a stranger, but who, as it happens, is planning to divorce his wife, it may be only *de dicto* concern for doing what is right that moves him to save her. They have also argued that the fact that a good person is motivated to do what she believes to be right, whatever that happens to be, does not preclude her from also being motivated non-derivatively by direct concern for another's welfare (Lillehammer 1997: 193). Furthermore, they argue, even if there were something fetishistic about being motivationally disposed to do the right thing, whatever it happens to be, the externalist has alternative ways of explaining the connection between moral judgment and motivation. For example, Copp (1997: 50–51) has suggested that an individual might be motivationally disposed to desire non-derivatively to do the very thing that

[1] For a more thorough reconstruction of Smith's argument, see Dreier (2000).

she judges it right to do rather than being disposed to do the right thing, whatever it happens to be. The person who judges that she ought to aid a stranger in distress would then be motivated not by a non-derivative desire to do the right thing but by a desire to do this particular right thing, namely aid the stranger.[2]

Likewise, Sigrun Svavarsdóttir (1999) has argued that Smith is mistaken in claiming that the only explanation of motivational shifting available to the externalist is one that appeals to a (*de dicto*) desire to do the right thing. She argues, though, that something close to the view Smith rejects provides the correct externalist account of moral motivation. The good person, she argues, should be understood as concerned with doing what is morally required or morally good where that *encompasses* what is honest, considerate, just, fair, and so on. It would be a mistake to suppose that the person thus motivationally disposed cares only about doing one thing, namely whatever she believes to be right, as Smith seems to suggest. It would also be a mistake to suppose that when the good person, so conceived, undertakes an act, she does so conceiving of that act only as the right thing to do. The externalist picture allows that the good person may often, for example, simply respond directly to the distressed stranger in need of aid. And that picture need not introduce a thought that alienates a person "from the ends at which morality properly aims," such as the alienated thought "it's the right thing to do."

Dreier (2000) finds Smith's fetishism argument largely successful against externalist appeals to the *de dicto* desire to do the right thing. But he argues that externalists can offer a different and more compelling model of moral motivation that answers the argument, one that appeals to second-order desires. In this model, what the good person desires is that, in all cases, if φ is the right thing to do, then she desires to do φ. In this model, although the second-order desire may play a causal role in a good person's coming to have his first-order motivations, say, to act for the welfare of his wife, once acquired, those motivations and even the desire to maintain those motivations may move him to act without any thought of their rightness.

Externalists insist that because not all persons who judge it right to φ are motivated to φ, and because wide variation exists in how people's moral judgments affect their feelings, deliberations, and actions, some conative state must effect the movement from judgment to motivation (Svavarsdóttir 1999: 161). Svavarsdóttir contends, for example, that the desire to be moral effects a "psychological transition" from a person's judgment that she ought to φ to her wanting to φ (1999: 201). Even if the desire, say, to aid a stranger in need, initially derives from a desire to do whatever morality requires, the desire to do that particular right act may come to motivate on its own, so that her desire to aid is not merely instrumental to her desire to do the right thing (Svavarsdóttir 1999: 205–206, 213–214).

[2] For a critique of this view, see Dreier (2000).

Direct vs deferred internalism

Internalists, pressed by externalist examples of apparent failures of motivation, have offered yet more qualified versions of internalism. Recall the example of the retired person who, after twenty years of working to relieve poverty, continues to judge (sincerely) that she morally ought to work to alleviate poverty but who is no longer motivated to do so. Such a person arguably does not suffer from practical irrationality, so Smith's conditional internalism might not resolve the problem. Some internalists have attempted to address this sort of case by suggesting that perhaps moral judgments need not be *directly* motivating to preserve a plausible necessary connection between moral judgment and motivation. Internalists might simply hold that necessarily, in cases in which a person judges that she morally ought to φ but is not directly motivated (at least somewhat) to φ, there are some relevantly connected moral judgments that do motivate the person making them. This view is sometimes referred to as *Deferred Internalism*.

> Deferred Internalism: Necessarily, if a person judges that she morally ought to φ, then she is either (at least somewhat) motivated to φ or some relevantly connected moral judgments are accompanied by motivation.
>
> (Björnsson et al. 2015: 9)

These relevantly connected moral judgments might be earlier judgments made the person herself; our retired aid worker was presumably motivated at earlier periods of her life by her moral judgment that she ought to work to alleviate poverty.[1] Alternatively, as Jon Tresan (2009a, 2009b) argues, internalists might adopt a version of what he calls *Communal Internalism*. (See also Blackburn 2001: 63.) According to the latter view, a person's beliefs and judgments only count as moral when, in her community, beliefs or judgments with that content are motivating. One motivation for the move to a form of Communal Internalism is the thought that the amoralist is only intelligible against the backdrop of a community in which moral judgments ordinarily motivate (Tresan 2006: 151; drawing on Foot 1978). In order for there to be moral beliefs,

> certain *practices* must exist (or ceremonies, rituals, habits, customs, what have you, the crucial thing being that they require conations). Once these practices are up and going the conative condition is satisfied and there

[1] A different way to develop an indirect form of internalism might appeal to second-order desires. An agent might not have a desires to φ but may have a desire that she desire to φ.

> may be moral beliefs. A community characterized by practices will contain individual members who don't participate in them, but pick up beliefs from those who do.
>
> (Tresan 2006: 150)

Amoralists might be such individual members, and if so, "they have moral beliefs not because there are no necessary conative conditions but because those conditions are *satisfied*" (150). Communal Internalism thus allows for *individual* amoralists and so offers a form of internalism that retains the core internalist idea, while being sufficiently weak to meet the amoralist challenge.

Externalists would no doubt object that even if the amoralist is only intelligible against the backdrop of a community in which moral judgments ordinarily motivate, the connection between the moral judgments of community members and motivation may nevertheless be contingent. How, then, are we to decide between Communal Internalism and externalism, which equally account for individual amoralists? For this, Tresan contends, we must consider the possibility of amoralist *communities*.

Imagine a community in which people reacted to what others did only when it affected them or their loved ones and in which people were not taught various moral rules, such as not to steal or murder (Foot 1978: 203–204). Imagine, too, that these people are in all other respects just like us. The question we must consider is whether they have moral beliefs or make moral judgments. Internalists will contend that they do not, that even if internalism doesn't seem plausible with respect to individual amoralists, it is surely the more plausible position with respect to communities.

Externalists will be unpersuaded by the thought experiment. They might concede that the people in our imagined community do not have moral beliefs or make moral judgments, while denying that this is for the reasons Communal Internalism offers. They might, for example, maintain that it is part of our concept of moral judgment that such judgments are universal in scope, and the people in question do not genuinely make judgments with this scope. Alternatively, externalists might argue that it is perfectly conceivable that a community of people might be taught moral rules but care about their infringement (and be motivated) only when they themselves are somehow affected. Perhaps such people are selfish or otherwise immoral, but they need not be guilty of insincerity or conceptual confusion. Externalists might thus continue to deny that it is a part of our concept of moral judgment that necessarily, moral judgments are at least somewhat motivating, even when motivation is at the level of communities.

De dicto vs de re internalism

Internalists have defended internalism as a thesis about what it is for a mental state or act to *count* as a moral judgment: for a mental state to count as a moral judgment, it must be accompanied by motivation.[1] This leaves open that the motivation involved may be conditional or unconditional, direct or indirect (as in the case of Communal Internalism).

As Tresan (2006: 143) characterizes internalism, "moral beliefs require motivational or affective states ("conations") ... *believing that x is right* requires *having a pro-attitude toward x*." But the claim that moral beliefs require conations, he says, admits of *de dicto* and *de re* readings.[2]

> De dicto Internalism: Necessarily, moral beliefs are accompanied by conations.
>
> De re Internalism: Moral beliefs are necessarily accompanied by conations.
>
> (Tresan 2006: 145)

De dicto internalism is a thesis about our *concept moral belief*; we conceive of moral beliefs as accompanied by conation. In contrast, *de re* internalism is a thesis about the beliefs that are moral beliefs in the actual world, that they are accompanied by motivation. *De dicto* internalism does not entail *de re* internalism, because de dicto internalism, as a thesis about our concept *moral belief*, leaves open the nature of moral beliefs. Tresan insists that *de dicto* internalism, in fact, tells us nothing about the nature of moral beliefs. "In itself, it tells us no more than Externalism" (Tresan 2006: 148).

Our intuitions about amoralists, namely, that they lack moral beliefs, "at best support de dicto Internalism," Tresan argues (2006: 148). Consider, he says, what would be involved in testing *de dicto* v. *de re* internalism. In testing the former, "we must consider whether there are possible states which (at one and the same world) are moral beliefs and are unaccompanied by the relevant conations" (149). In contrast, in testing the latter thesis, "we would consider whether conations are necessary for the existence of *those very things* which, in the actual world, are moral beliefs" (149). But when we consider the amoralist challenge, and have the intuition that the amoralist at world W lacks moral beliefs, we are not considering whether he might have states that are not

[1] Björnsson et al. (2015: 11) call this "non-constitutional internalism."

[2] The distinction Tresan draws is unrelated to Smith's *de re*/*de dicto* distinction discussed earlier.

moral beliefs at W but are moral beliefs in the actual world. Rather, we are considering whether at W, he has states that are moral beliefs.

Externalists would likely reject even *de dicto* internalism, as weak as it might seem. Their position in raising the amoralist challenge is, after all, that the amoralist is *conceivable*. Moreover, as we have seen, they have a story to tell about why at least certain amoralist communities might be inconceivable that does not favor internalism.

Motivational internalism as an empirical thesis

Although internalism, in its various forms, has been treated as an a priori or conceptual thesis about moral judgment, some philosophers have recently treated it as an empirical thesis (or, in Tresan's terminology, have treated internalism as a *de re* view), or at least as a thesis on which empirical evidence might be brought to bear. Some have argued that empirical evidence counts against internalism, and others, that it supports internalism.

Adina Roskies targets the view that she calls "motive-internalism," which she characterizes as the view that "moral belief entails motivation" or that "motivation is intrinsic to, or a necessary component of moral belief or judgment" (2003: 51–52). According to Roskies, the motive-internalist faces a dilemma: either the internalist thesis is a weak thesis about the nature of moral belief and so is philosophically uninteresting, or it is strong enough to be philosophically interesting, but it is false.

Smith's internalism, she argues, provides an example of an internalist thesis that falls on the first horn of the dilemma. Recall that on his view, an agent who judges that it is right to φ in circumstances C is either motivated to φ in C or is practically irrational. Roskies argues that this view requires an account of what it is to be practically rational. But if being practically rational is a matter of desiring to act as one judges best, then Smith's internalism is trivially true (53). It amounts to a definitional claim about practical rationality rather than a strong claim about a necessary connection between moral judgments and motivation. Roskies also considers what she calls the internalist thesis that "Usually/Normally, if an agent believes that it is right to φ in circumstances C, then he is motivated to φ in C," arguing that it, too, is too weak to be philosophically interesting (53–55).[1] In any case, given what she treats as a sufficiently strong version of internalism in her discussion of the other horn of the dilemma, she would presumably reject any form of conditional internalism as too weak.

Roskies characterizes the form of internalism that she would consider sufficiently strong to be philosophically interesting as follows, referring to it as *substantive internalism*:

> SI: If an agent believes that it is right to φ in C, then he is motivated to φ in C.

[1] Roskies does not attribute the view to particular philosophers, but perhaps she has in mind

(2003: 55)

But this thesis, she contends, is "empirically false," and so falls on the second horn of the dilemma (55). As evidence of this, she examines research on patients with damage to the ventromedial (VM) prefrontal cortex, which is "anatomically connected to a wide variety of brain areas, including those associated with perception, reasoning, declarative knowledge, and with emotion and visceral control" (55). "VM patients" – those with so-called "acquired sociopathy" – appear in psychological testing to be normal in intelligence and reasoning ability, but they have difficulty acting as they judge appropriate (56). According to researchers, although VM patients make moral claims like those of "normals," they fail reliably to act as normals do; they also "seem to lack appropriate motivational and emotional responses" when it comes to moral matters, though they retain other ordinary motivations, like seeking food and company (57). The evidence for thinking that VM patients are not motivated by their moral beliefs or judgments comes from examining measurable skin-conductive response (SCR), the presence of which Roskies treats as evidence of motivation and the absence of which she treats as evidence of the absence of motivation. As compared with normals, when presented with "emotionally-charged or value-laden stimuli," VM patients "do not generally produce SCRs" (57). Thus, although they make sincere moral judgments, they fail to be motivated (59).[2] VM patients are, she contends, a "walking counterexample" to internalism.

Critics have offered a number of arguments against Roskies' position. Kennett and Fine (2007: 182) observe that in order to support her position, Roskies must show both that VM patients indeed make the relevant moral judgments and that the best explanation of their failure to act in accordance with their moral judgments is a deficit in moral motivation. But Roskies appears to rely on just one case study as evidence that VM patients make the relevant moral judgments – that of a VM patient referred to as EVR. The tests of moral judgment conducted on EVR, moreover, all involved third-personal reasoning and third-personal moral dilemmas. They thus did not involve the first-personal, "I ought" judgments that bear on internalism. Is EVR even capable of first-personal moral reasoning? Does he recognize when he is in a situation covered by his third-personal moral judgment and transition to the relevant "I ought" judgment? Some evidence at least suggests, they argue, that VM patients may not appropriately connect third-personal and first-personal judgments, and moral reasoning has not been sufficiently studied in such patients. Turning to the second point, Kennett and Fine claim that information about EVR suggests that his behavior is better explained by a general impairment in decision-making rather than a failure of moral motivation

---

something like the view defended by Dreier (1990).

[2] Roskies acknowledges that some versions of internalism may be consistent with the data on VM patients, though versions that she herself would consider either problematic or insufficiently developed (2003: 62–63).

(184).[3] They also dispute Roskies' claim that the situations in which VM patients failed to show SCRs were ethically charged and her assumption that the absence of SCR reliably indicates the absence of motivation (187–188).

Schroeder et al. (2010: 95) report that research indicates that psychopaths show "diminished capacity to distinguish moral from conventional violations"; as a consequence, some have concluded that psychopaths have impaired moral concepts (2010: 96, citing Nichols 2004). VM patients who suffer their injuries later in life, it has been argued, differ from psychopaths in that they do not exhibit moral deficits. Rather, "their deficits in non-moral aspects of life merely manifest occasionally in moral situations." The extant data on VM patients indicates that whereas those who were injured early in life exhibit sociopathic behavior (including violent behavior), those who were injured later in life do not. What might explain this difference remains unclear. Are the latter patients not violent out of habit, say, or because their moral judgments are to some degree motivating (98)? More research is obviously needed. Roskies herself allows that the evidence regarding VM patients is inconclusive (2007: 205).[4]

Eggers (2015: 95) argues that the most serious difficulty for Roskies' argument is that her reasons for believing in a strong connection between SCRs and motivation aren't good enough. (See also Leary forthcoming.) The absence of SCRs has been reliability correlated with cases in which persons have failed to act in accordance with their judgments, but this would indicate only that they are reliable indicators of a person's *overriding* motivation (the one that issues in action). He argues further that "it is hard to see how we could ever produce evidence that certain physiological phenomena ... are strictly correlated with motivation, other than inferring this from the way these phenomena correlate with human behavior" (96). But behavior bears only on overriding motivation, and so the absence of behavior doesn't refute internalism.

Even assuming that the evidence about VM patients shows what Roskies suggests it does, various forms of conditional internalism and deferred internalism might well be able to allow for VM patients. Roskies might regard such views as insufficiently strong to be of real philosophical interest, but internalists would insist that their views are informative about our concepts *moral belief* and *moral judgment*. What remains unclear is the bearing of the empirical evidence that Roskies cites on internalism of any of the forms we have considered. After all, as advanced by its proponents, internalism is an a priori or conceptual thesis.

Jesse Prinz (2015: 61) has argued, contrary to the way internalists ordinarily understand their view, that internalism can be understood as a *psychological* thesis. So

[3] Kennett and Fine (2007: 15–186) go on to examine critically Roskies' efforts to rule out alternative explanations related to the one that they offer.

[4] Other criticisms of Roskies' arguments include those offered by Prinz (2015: 17), who contends that evidence that purports to show that VM patients make moral judgments without feeling emotion "is not decisive," by Cholbi (2006), who offers reasons to doubt that VM patients have moral beliefs, and by Gerrans and Kennett (2010), who argue that VM patients have impaired moral agency and so do not make genuine moral judgments. Roskies (2006, 2007) offers various replies to her critics.

understood, he contends, empirical evidence actually *supports* internalism. Prinz offers a number of arguments for this claim. Perhaps the most interesting of these appeals to a view called *sentimentalism*.[5]

1. Moral judgments consist of emotional attitudes.

2. Emotional attitudes are motivating.

3. Therefore, moral judgments are motivating (70).

According to Prinz, sentimentalism is the view that "moral judgments consist of feelings directed at whatever it is that we moralize" (70). So premise 1 is just a statement of sentimentalism. This view, he argues, supports a number of empirical predictions, and the evidence confirms these predictions. Various studies provide evidence "that people enter into emotional states when they make moral judgments" (71), that induced emotions have an impact on moral judgments, with different emotions having different effects (72), and that people with different emotional dispositions differ in their moral judgments. Thus, studies have found that when happiness is induced, people tend to make more positive moral judgments, and when disgust is induced, they tend to judge scenarios involving moral wrongness more harshly. Persons with Huntington's disease have deficits in disgust and exhibit sexual deviancy, whereas "psychopaths, who have deficits in several negative emotions, but not disgust, show insensitivity to crimes against persons, but are not known for sexual deviancy" (73). Evidence like this, according to Prinz, adds "support to the claim that emotions are components of moral judgments. Emotions occur when people make moral judgments, they are used as information when reporting strength of moral attitudes, and emotional deficits lead to corresponding deficits in moral sensitivity" (73). The link between emotion and action in premise 2, he says, is supported by decades of research. The argument thus provides strong support for the internalist idea that moral judgments are motivating.

Prinz's ways of talking about sentimentalism are not entirely consistent. He sometimes says that emotions are "components" of moral judgments, and sometimes that moral judgments "consist of" emotional attitudes, and these ideas are not equivalent. But even if emotional attitudes are implicated in moral judgments – that is to say, even if a person who judges that she ought to φ experiences and evinces an emotional attitude – it is doubtful that moral judgments themselves consist of emotional attitudes. What's more, the evidence to which Prinz appeals does not obviously support sentimentalism. It may instead simply show that emotions accompany moral judgment.

[5] For discussion of the other arguments, see Prinz (2015) and Rosati (2016).

Unconditional internalism revisited

We have considered various qualifications of Unconditional Internalism that have been urged by internalists. Eggers (2015) has recently argued, as against the trend of qualifying internalism, that internalists should stick with unconditional motivational internalism (UMI). UMI, he claims, is in fact a very weak thesis, and once this is appreciated, internalists should recognize that there is no need to qualify it.

Eggers argues that the cases urged against internalism either offer inappropriate counterexamples or are simply inconclusive (2015: 88). Examples offered to show that a person can make a moral judgment while lacking moral motivation tend to involve that person's failure to *act* in accordance with her moral judgment. But because UMI says only that necessarily, a person who judges that she morally ought to φ is somewhat motivated to φ, a person's failure to act can at most be evidence that she lacks *overriding* motivation to φ, not that she lacks any motivation at all to φ. A person might, after all, have conflicting desires, one of which proves to be stronger and which overrides the other, leading her to act. As long as the imagined cases said to count against internalism allow for more than one motivation, "the agent's behavior will tell us nothing about what is crucial for assessing UMI: whether any moral motivation was present that was too weak to produce action in accordance with the moral judgment" (89).

It does not help to add, as some critics of internalism do, that the person reports having no motivation (89). Although an agent may be a reliable judge as to his overriding motivation, we have no reason to assume that he is reliable when it comes to weak motivation. The agent may not even be aware of motivational conflict, as not all such conflict is conscious (90). Following Finlay (2004: 209), Eggers concludes that "one of the problems of the internalism/externalism debate is the notorious difficulty of proving the absence of (relevant) motivation" (90).

What we need, he suggests, is a reliable test case – one that "allows us better to distinguish between absent motivation, on the one hand, and overridden motivation, on the other" (90). To eliminate possible bias toward the externalist, the test case must also, insofar as possible, eliminate the possibility of conflicting motivation. The case must be such that the agent can "perform the morally required action almost effortlessly" so as to eliminate a possible conflict between a motivation to act in accordance with her moral judgment and a motivation to avoid having to exert any effort. And the action must be indirectly related to the agent's moral judgment so that she is not the addressee of the judgment and so is not likely to experience a conflict between moral motivation and some non-moral motivation. Only if the agent fails to act in a test case

that meets these conditions will it be implausible for the internalist to explain away the agent's failure to act as due to some conflicting non-moral motivation. In keeping with the requirements of the test case, Eggers suggests that we amend the definition of UMI as follows:

> *Unconditional motivation internalism** (UMI*): necessarily, if a person judges that it is morally wrong to φ, then she is, at least to some extent, motivated to refrain from φ-ing *and/or to keep others from φ-ing*".
>
> (92, emphasis added)

Eggers then asks us to imagine an agent who believes that lying is morally wrong and must choose between two possible states of the world. In state A, people are honest with one another, while in B, which is in all other respects the same, people regularly lie to one another. Imagine that apart from the agent's belief that lying is wrong, he has no other incentives for choosing one world over the other. He has no general desire to do what he judges morally right and avoid what he judges morally wrong. He has no personal attachments to the people in either world. He will not suffer any reputational consequences of his choice. And he will himself live in neither world. Whereas the externalist must claim that the agent in this example will be unable to choose because he will be indifferent as between worlds, the internalist must claim (more plausibly, Eggers thinks) that the agent will choose state A because he judges it morally better.

The test case, as posed, may seem to treat internalism as an empirical thesis. But as we have already seen, Eggers rejects efforts to refute UMI as an empirical psychological claim. Still, he suggests, empirical evidence that is *linguistic* rather than psychological may bear on its truth as a conceptual claim. In contrast to psychological evidence, where we cannot be sure (for reasons we have seen) that the evidence bears on the truth of internalism, linguistic evidence can tell us about our concepts and so is relevant to understanding our concept of moral judgment (97, 98–99). Eggers does not claim that we can identify the features of our moral concepts just by studying ordinary speakers' linguistic habits and assumptions. Rather we identify those features by "critically reconstructing" our moral concepts. Eggers observes that although there has been little study of the relevant linguistic evidence with respect to UMI, his test case suggests "an empirical approach for identifying the relevant aspects of our concept of moral judgment" (98).

Eggers reports the results of experiments he conducted at a number of universities in which he gathered ordinary German speakers' views about the test case and variations on it. According to Eggers, these experiments suggested that a large majority of ordinary speakers accept UMI and so hold an internalist view of moral judgment (101), which speaks "in favour of UMI as a conceptual claim" (103). Nevertheless, he acknowledges that the results do not allow us to conclude that UMI is true (104). After all, the results leave us with conflicting evidence (large numbers of ordinary German

speakers apparently did *not* hold an internalist view of moral judgments), so the question remains as to whether there is a way to decide as between the conceptual claims of the internalist and the externalist. Ultimately, Eggers concludes that the results show only that "UMI does capture a feature of most ordinary speakers' concept of moral judgment – capture it, that is better than externalism or weaker versions of internalism" (106).

Externalists might well argue that Eggers' results can be explained away. Perhaps despite the various constraints built into the test case, subjects were not able to keep these firmly in mind in making their responses. Their inability to do so might well be explicable. For one thing, we tend to expect that people are ordinarily concerned about the effects of actions on other people, even in the absence of personal attachments. If externalists can successfully explain away the responses of those subjects who seemed to accept an internalist view of moral judgments, then the externalist would seem to be well positioned to reject linguistic evidence that apparently supports internalism. Without more, then, it would appear that Eggers' experiments based on his test case leave the debate between internalists and externalists unsettled in more than one way.

Conclusion

Should internalists and externalists be content to leave the debate where Eggers seems ready to leave it? Perhaps people simply have different concepts of moral judgment, some of them being internalist and some externalist. This will no doubt strike many – at least, parties to the debate – as unsatisfying, particularly those who think that the truth of internalism would have broader implications for theory choice in metaethics. Still, as we have seen, the various efforts to refine internalism so as to meet the amoralist challenge narrow the gap between internalism and externalism considerably. Unless the internalist can establish convincingly that one or another refinement preserves a plausible necessary connection between moral judgment and motivation, the externalist will insist that the internalist has been too long in search of a connection that does not exist.

References

Blackburn, S. (2001) *Ruling Passions*, Oxford: Clarendon Press.

Björnsson, G. (2002) "How Emotivism Survives Immoralists, Irrationality, and Depression," *Southern Journal of Philosophy* 40: 327–344.

Björnsson, G., Strandberg, C., Ollinder, R., Eriksson, J., and Björklund, F. (eds.) (2015) *Motivational Internalism*, Oxford: Oxford University Press.

Brink, D. (1989) *Moral Realism and the Foundations of Ethics*, Cambridge: Cambridge University Press.

———. (1997) "Moral Motivation," *Ethics* 108: 4–32.

Cholbi, M. (2006) "Belief Attribution and the Falsification of Motive Internalism," *Philosophical Psychology* 19: 607–616.

Copp, D. (1997) "Belief, Reason, and Motivation: Michael Smith's *The Moral Problem*," *Ethics* 108: 33–54.

Darwall, S. (1983) *Impartial Reason*, Ithaca, NY: Cornell University Press.

Dreier, J. (1990) "Internalism and Speaker Relativism," *Ethics* 101: 6–26.

———. (2000) "Dispositions and Fetishes: Externalist Models of Moral Motivation," *Philosophy and Phenomenological Research* 61: 619–638.

Eggers, D. (2015) "Unconditional Motivational Internalism and Hume's Lesson," in G. Björnsson, C. Strandberg, R. Ollinder, J. Eriksson, and F. Björklund (eds.), *Motivational Internalism*, Oxford: Oxford University Press, 85–107.

Falk, W. (1952) "'Ought' and Motivation," in W. Sellars and J. Hospers (eds.), *Readings in Ethical Theory*, New York: Appleton – Century-Crofts.

Finlay, S. (2004) "The Conversational Practicality of Value Judgment," *Journal of Ethics* 8: 205–223.

Foot, P. (1978) "Approval and Disapproval," in *Virtues and Vices*, Berkeley and Los Angeles: University of California Press.

Frankena, W. (1976) "Obligation and Motivation in Recent Moral Philosophy," in K. Goodpaster (ed.), *Perspectives on Morality: Essays of William Frankena*, Notre Dame, IN: Notre Dame University Press.

Gerrans, P. and Kennett, J. (2010) "Neurosentimentalism and Moral Agency," *Mind* 119: 585–614.

Hare, R. (1963) *Freedom and Reason*, Oxford: Oxford University Press.

Kennett, J. and Fine, C. (2007) "Internalism and the Evidence from Psychopaths and 'Acquired Sociopaths,'" in W. Sinnott-Armstrong (ed.), *Moral Psychology Vol. 3: The Neuroscience of Morality*, Cambridge, MA: MIT Press, 173–190.

Leary, Stephanie. (forthcoming) "Defending Internalists from Acquired Sociopaths," *Philosophical Psychology.*

Lillehammer, H. (1997) "Smith on Moral Fetishism," *Analysis* 57: 187–195.

McDowell, J. (1978) "Are Moral Requirements Hypothetical Imperatives?" *Proceedings of the Aristotelian Society* 52: 13–42.

———. (1979) "Virtue and Reason," *Monist* 62: 331–350.

Nichols, S. (2004) Sentimental Rules: On the Natural Foundations of Moral Judgment*, New York: Oxford University Press.*

Parfit, D. (2011) *On What Matters*, Vol. 2, New York: Oxford University Press.

Prinz, J. (2015) "An Empirical Case for Motivational Internalism," in G. Björnsson, C. Strandberg, R. Ollinder, J. Eriksson, and F. Björklund (eds.), *Motivational Internalism*, Oxford: Oxford University Press, 61–84.

Rosati, C. (1996) "Internalism and the Good for a Person," *Ethics* 106: 297–326.

———. (2016) "Moral Motivation," *Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/entries/moral-motivation/.

Roskies, A. (2003) "Are Ethical Judgments Intrinsically Motivational? Lessons from 'Acquired Sociopathy,'" *Philosophical Psychology* 16: 51–66.

———. (2006) "Patients with Ventrodemial Frontal Damage Have Moral Beliefs," *Philosophical Psychology* 19: 617–627.

———. (2007) "Internalism and the Evidence from Pathology," in W. Sinnott-Armstrong (ed.), *Moral Psychology Vol 3*: *The Neuroscience of Morality*, Cambridge, MA: MIT Press, 191–206.

Sarch, A. (2010) "Internalism about a Person's Good: Don't Believe it," *Philosophical Studies* XXX.

Schroeder, M. (2007) *Slaves of the Passions*, Oxford: Oxford University Press.

Schroeder, T., Roskies, A., and Nichols, S. (2010) "Moral Motivation," in J. Doris (ed.), *The Moral Psychology Handbook*, Oxford: Oxford University Press, 72–110.

Shafer-Landau, R. (1998) "Moral Motivation and Moral Judgment," *Philosophical Quarterly* 48: 353–358.

———. (2000) "A Defence of Motivational Externalism," *Philosophical Studies* 97: 267–291.

———. (2003) *Moral Realism: A Defence*, Oxford: Clarendon Press.

Smith, M. (1989) "Dispositional Theories of Value," *Proceedings of the Aristotelian Society, Supplementary Volumes* 63: 89–111.

———. (1994) *The Moral Problem*, Malden, MA: Blackwell.

Svavarsdóttir, S. (1999) "Moral Cognitivism and Motivation," *Philosophical Review* 108: 161–219.

———. (2006) "How do Moral Judgments Motivate?" in J. Dreier (ed.), *Contemporary Debates in Moral Theory*, Malden, MA: Blackwell Publishing, 163–181.

Timmons, M. (1999) *Morality Without Foundations*, Oxford: Oxford University Press.
Tresan, J. (2006) "De Dicto Internalist Cognitivism," *Noûs* 40: 143–165.
———. (2009a) "Metaethical Internalism: Another Neglected Distinction," *Journal of Ethics* 13: 51–72.
———. (2009b) "The Challenge of Communal Internalism," *The Journal Of Value Inquiry* 43: 179–199.
Williams, B. (1981) "Internal and External Reasons," in *Moral Luck*, Cambridge: Cambridge University Press.
Zangwill, N. (2015) "Motivational Externalism: Formulation, Methodology, Rationality, and Indifference," in G. Björnsson, C. Strandberg, R. Ollinder, J. Eriksson, and F. Björklund (eds.), *Motivational Internalism*, New York: Oxford University Press, 44–60.

17. Emotions in practical reasoning[1]

[1] Let me thank co-editor Kurt Sylvan and students in my 2016 and 2018 seminars for very helpful comments on earlier drafts of this chapter.

Patricia Greenspan

Emotional sources of action are typically treated in contrast to practical reasoning, even by contemporary authors who argue for their rationality in certain cases (e.g., Arpaly 2003). My own work on emotions has focused instead on showing how they can play a valuable role *within* practical reasoning, both in directing attention to reasons for action and in augmenting their normative force.

My aim in approaching this subject was not to provide a comprehensive theory of the nature of emotions but rather to understand something about their rational role as a supplement to – or sometimes even a substitute for – judgment. But I first needed to argue against the then-dominant view of emotions as themselves amounting to judgments (see, e.g., Solomon 1993): evaluative judgments, which I take to include beliefs about one's reasons for action, of the sort featured in standard conceptions of practical reasoning.

At the same time, I wanted to exhibit a relation of emotions to evaluative judgments that would allow for similar sorts of justificatory assessment – what I thought of as the "logic" of emotion (see Greenspan 1980). What I intended was not simply to weaken the element of judgment in the dominant view to make it less vulnerable to objection. Nor did the point of preserving the element of evaluation depend on accepting the claim underlying the dominant view, that emotion-types could be individuated only on that basis, with no reference to qualities of affect. Rather, what I had in mind was to use the idea of an evaluative content of emotion to connect emotions to reasons.

In what follows, I hope to clarify my approach to the subject, correct some features of my earlier presentation, and then connect it to some later work I did on practical reasons, arguing that standard conceptions of them overstate their normative force. I began the job of connecting my two projects in a recent piece on emotions as reinforcing moral reasons (Greenspan 2012), but I think the role of emotions is more likely to be decisive in justifying action on nonmoral reasons, even in some cases where warranted judgments fall short. So that will be my main focus here.

1 Affect as evaluative

I think of emotional affect, or the feeling aspect of emotion, as positive or negative in valence, with correspondingly positive or negative evaluative content that can be considered in isolation for purposes of justificatory inquiry. Thus, feelings of joy give a positive evaluation of something; feelings of fear involve the thought of something as dangerous. Other, more complex emotions may involve several such pairs of affect and evaluation, some positive, some negative: love and anger, for instance, may both include negative evaluations of an unfulfilled desire, for closeness or for retribution, though the primary evaluative content of love – evaluating the object of love, the person or thing loved – is of course positive.

Note that I make a distinction here between object and content. In Greenspan 1988, I referred to the evaluative object of emotional affect as the "internal object" of emotion, and while the term still seems apt, it apparently confused many readers. So I now speak of emotion as having an evaluative *content* (see Greenspan 2019). Both notions, object and content, can be said to capture what a feeling is about; one might think of them as analogous to sense and reference. Fear of a dog, for instance, refers to the dog and signifies that it is dangerous. Its content can be represented as an evaluative proposition, whereas its object is simply the dog.

The point of this sort of analysis – I think of it as a way of "parsing" emotions – is to let us extract something propositional from an emotion to contrast with judgment in considering justificatory questions, questions about the rational justification of emotion and of action motivated by it: whether holding that content in mind with some degree of affect is warranted by the available reasons and whether affect appropriately directed toward it provides a reason for action beyond what is provided by warranted belief.

While I speak of emotions, thus construed, as having affective and evaluative elements or components, this is not meant to imply that the two are separable in fact or that in undergoing an emotion one independently entertains a corresponding evaluative thought. Affect does not simply *accompany* an evaluation, as suggested by Aristotle's definition of anger, which can be seen as a source for this general sort of account:

> Anger may be defined as an impulse, accompanied by pain, to a conspicuous revenge for a conspicuous slight directed without justification towards what concerns oneself or towards what concerns one's friends.
>
> (*Rhetoric* 1941: 137830:2–5)

Instead, as I put it in Greenspan 2003: "Affect evaluates." An evaluative proposition expresses what a feeling is trying to tell us, as it were. Occasionally, we have to hunt

for this, inquiring into a nagging feeling by reviewing the day's events, say, to *figure out* what its content is and hence what emotion it amounts to. For that matter, its evaluative content need not be articulable by the subject of emotion – as is evident in the case of animals and infants, who lack the conceptual equipment for other forms of expression but still are capable of pro or con reactions.[1] But we theorists can isolate evaluative content for purposes of considering justificatory issues.

So despite my focus on evaluative content, the role I attribute to affect in modifying the view of emotions as evaluative judgments actually results in something more like the competing, "perceptual" view in the literature, which in its classic form takes emotions as perceptions of value (see, e.g., Sousa 1987). However, I would not take the implied analogy to sense-perception seriously; in particular, I would reject the suggestion that emotional affect necessarily evaluates something the subject sees as a real property of the object of emotion. Fantasy results in genuine emotions, sometimes in response to evaluations the subject does not accept. Consider my enjoyment, say, at the thought of some dire fate befalling a bully. I might assign value to some lesser form of punishment, but that is not all I am reacting to in fantasy; nor does my reaction imply some momentary delusion about what I really value.

The affective element of emotion may also, of course, be relevant to justificatory issues. Most obviously, there are disproportionate emotional reactions such as extreme anger at a minor slight. But I see no prospect of giving a systematic account of what justifies a given intensity of affect, beyond requiring some rough proportionality to the strength of the reasons for emotion. So my focus has been on the evaluative element. Moreover, my account makes no attempt to capture specific features of affect – the detailed phenomenology of emotional experience – though I would not dismiss its importance. Instead, I simply classify affect by its positive or negative valence. Initially, in fact, in Greenspan 1988 I used some rather regimented terminology for this – "comfort" or "discomfort" – in the hope of imposing some artificial order on an area that seemed by nature rather messy. My emphasis was on emotional discomfort, as a broad and familiar term for negative affect. "Comfort" was admittedly unidiomatic for many cases of positive affect, but I chose it just for its verbal contrast to "discomfort." I now use more varied language, but for my purposes, what matters is that emotional discomfort both evaluates something negatively – has a negative evaluative content – and *feels bad*. It therefore can provide a reason – both a motive and a normative reason – to act to falsify the evaluation, in the sense of making it inapplicable by changing the situation. Other things being equal, emotional discomfort is a state of an agent that warrants change. That, in a nutshell, is how I take emotions to figure in practical reasoning.

[1] For an account of how infant emotions might develop complex cognitive content, see Greenspan 2010.

2 Warrant for emotion

In Greenspan 1988, in order to establish emotional bases for sound practical reasoning, I first addressed the question of whether and when a given emotion is itself warranted – justified by the available reasons, or appropriate in a rational (as distinct from a social or moral) sense. Hume (1978: 415–416, 458ff.) famously denies that either the passions or the actions they motivate are ever reasonable or unreasonable, but both claims can be contested. With respect to the reasonableness of emotion – emotional appropriateness, in my terms – content provides something that can be said to have representational quality, insofar as it can reflect or fail to reflect one's available reasons. Anger at a slight to a friend, say, evaluates some action as expressing undeservedly low regard for her and is appropriate only if one has no reason to think that she has done something to merit low regard or that the apparent slight is really just a bit of playful banter from another friend of hers.

Earlier I did not specify "available" reasons, by which I mean reasons the subject of emotion has access to. But later work on reasons and rationality has led me to make some further distinctions. By emotional "appropriateness," I had in mind, not quite an analogue of truth but rather something more like epistemic *justification*: warrant for the emotion, amounting to warrant for holding its content in mind. Thus, "rational" appropriateness is appropriateness to reasons. Since I now understand reasons in an objective sense, as facts about the situation rather than mental states, one might expect an emotion to count as appropriate as long as there are adequate reasons for it, whether or not one has any way of knowing the reasons. But understanding appropriate emotions in terms of rational justification rather than correctness introduces an element of subjectivity: one must in some sense possess the reasons that justify an emotion.

This does not entail ability to spell out the reasons. In Greenspan 1988, ch. 2, working with a case of suspicion, I argued that an emotion may be warranted even where the corresponding judgment is not – where one lacks sufficient evidence for believing someone untrustworthy, say, but responds emotionally to features of his manner that yield sufficient reason to keep that thought in mind. My suggestion was that, besides the evidential considerations that justify a judgment, general practical considerations also count among the reasons for attention to the corresponding evaluation, as secured by affect.[1] In effect, they can set a lower standard of evidence. Often the point is to allow for rapid response – to an object of fear, say (and I take suspicion to be a variant of fear), where one often cannot postpone a decision until all the evidence is

[1] Since I wrote, some epistemologists (e.g., Stanley 2005) have argued that non-evidential practical

in. But there are also considerations of general importance that make it reasonable to register an evaluation in affect without adequate grounds for an all-things-considered judgment.

I summed up such practical considerations as "general adaptiveness," but besides suggesting evolutionary adaptiveness, which I did not intend, that term may be too narrow in its reference to consequences. Consider again the case of anger at a slight to a friend: it might be thought that friendship itself entails somewhat heightened sensitivity to signs of a slight to a friend. But, of course, there are limits. One should not leap to extreme anger, but perhaps just be taken aback, by a remark that seems insulting. Consider jealous anger in this context: a twinge of it may be justified by a romantic relationship just in reaction to signs of the loved one's interest in someone else, though jealous *rage* is a prime example of an irrational emotion. Besides disproportionate affect, moreover, such reactions may involve misinterpreting clearly innocuous cues. Evidence for the emotional evaluation still is the central issue in assessing appropriateness, but my suggestion is that other factors can affect the standard of evidence. This is where moral factors might play a role, though indirectly – with the standard of evidence affected by an obligation to defend others from racial slights, for instance.

On similar grounds, and with similar limitations, I had argued in my earliest piece on emotion (Greenspan 1980) that emotional ambivalence, in the sense of conflicting emotions, may sometimes be rationally justified, in contrast to holding both of two contrary all-things-considered evaluative judgments. There also would seem to be enough in my account of appropriate emotion to support an argument for counting some cases of overt action on emotion (beyond the mental act of attention to an evaluation) as rationally justified in the absence of justification for the corresponding judgment – even some cases where explicit reasoning yields a conflicting judgment. An example might be hesitating to do business with the object of warranted suspicion, though one intellectually dismisses the features of his manner that prompt suspicion and con-

---

considerations such as how much is at stake in a given case can also affect whether a belief counts as knowledge – and hence, presumably, a justified judgment. But besides being disputed (see, e.g., Roeber 2016), this "pragmatic encroachment" view resists easy comparison with what I had in mind for emotion. Most obviously, the main case cited in support of pragmatic encroachment involves raising, rather than lowering, the normal standard of evidence. Indeed, in my suspicion case, I even allow an emotion to be justified by unspecified cues: evidence one can cite only vaguely, at best. But further, what I took to affect the standard of evidence for emotion was the *general* practical adaptiveness of holding a certain kind of content in mind, whereas in pragmatic encroachment, "what is at stake" refers to features of the particular case and is taken to justify an ought-judgment requiring specific action – in contrast to the evaluation in my suspicion case of a person as untrustworthy, which if held as a judgment would seem to have broader implications for action than would be reasonable on a lower standard of evidence. In the first instance, the act I consider justified in the case involves simply holding an evaluation in mind – which in turn, in the absence of countervailing reasons, warrants some caution in extending trust but not a charge of dishonesty or the like. For those who are skeptical of pragmatic encroachment even for emotions, my argument need only be that there is a lower standard of evidence for attention to an evaluative content than for endorsing a judgment to the same effect.

cludes – even rightly concludes – that one really has no adequate reason to judge him untrustworthy.

3 Reconstructing emotional reasoning

What I ultimately had in mind was to question the usual dichotomy between emotion and reasoned judgment as sources of action. I wanted to say that emotions can *provide* genuine reasons for action as well as rationally responding to them. To that end, I suggested an analogue of the practical syllogism framed in terms of emotion (Greenspan 1988, ch. 6, 2012). It was put in the first person, with a major premise attributing to a subject of anger an unfulfilled desire for retribution, by representing the emotional evaluation in question as an act-requirement: I am uncomfortable at the thought that I ought to get back at X.

But this may have misled many readers. It was intended only as part of a reconstruction of implicit practical reasoning, not as a description of how a real-life subject of emotion would be likely to think. Something similar might be said of the traditional practical syllogism: it serves as a way of making explicit the steps behind what in many cases is a quick leap to action on the basis of stored general knowledge (of a certain food as good for you, say, as in Aristotle's example) and immediate perceptual cues. Though we sometimes do deliberate step-by-step about such matters, the syllogism also can be taken as "unpacking" what is involved in just reaching into the refrigerator and grabbing a healthy snack – some leftover chicken, perhaps. In the case of action from emotion, explicit step-by-step reasoning – deliberation about one's state of discomfort and how to relieve it – is even less likely. After all, the main point of the affective state is to direct attention elsewhere – to its evaluative content, in this case an act-requirement that the subject has yet to satisfy. When she does act, she acts in light of the requirement, but – I want to say – with normative pressure to act added by the need to alleviate her discomfort.

Psychologists and cognitive scientists nowadays make a distinction between System I and System II processes, with the former understood as automatic and effortless, the latter as involving conscious steps. Emotions are often thought to be located squarely in System I and reasoning in II (see, e.g., Haidt 2001) – indeed, reasoning sometimes seems to be *defined* as involving conscious steps – but I think too sharp a distinction is a mistake. In my example of relatively automatic practical reasoning – choosing a healthy snack from the refrigerator – one acts in light of the awareness that the food in question is healthy, perhaps in contrast to some other possible choices. Perhaps there are other *healthy* choices, for that matter – the practical syllogism need not pick out a *necessary* means to one's ends – and opting for one of them may or may not be totally

unreflective, even if quick. Similarly for action prompted by emotion: there may be other ways of alleviating discomfort at an unsatisfied act-requirement, but acting to satisfy the requirement is the natural response. In short, there are practical inferences, and acts rationalized by them, without conscious steps.

I should note that I do not take the view associated with Hume that action must always be motivated by an emotion (in his terms, a passion). Awareness of an all-things-considered reason for action – holding an ought-judgment – can be enough. But what emotional discomfort adds to this is a further reason that is both self-regarding and a reason to act *soon*. It is an unpleasant state that seems likely to continue unless and until one changes the situation. I think of this as emotional pressure, a reason against *postponing* action. In Greenspan 1988, combining this with my treatment of emotional appropriateness, I argued that emotional discomfort can sometimes even tip the balance in favor of action on an evaluation that is not thought to be warranted as a judgment. In the case of suspicion, one might reasonably refrain from buying what someone is selling in light of the possibility that the uneasiness one feels is keyed to some signs of bad character or intentions that one cannot specify sufficiently to ground a judgment of untrustworthiness. Nor need one have a record of reliability in such matters that would support taking one's reaction as itself constituting adequate evidence for the judgment.

A more extreme response – accusing the man of dishonesty, say – clearly would not be justified in such a case. My argument is not meant to support "trusting your feelings," any more than salesmen, without restraint. Otherwise, allowing for the rationality of emotional ambivalence, as I did in Greenspan 1980, would mean endorsing action at cross-purposes. Reasonable expression of emotions in action often requires careful control, but here again I made no attempt to give an account of matters of degree. I also had little to say about action *expressive* of emotion but not instrumental to any further practical end on the order of self-protection in the suspicion case. In cases like crying out of sadness or jumping for joy, emotion can be said to provide a reason for action – or even *the* reason, a pressuring reason, and sometimes perhaps even a normative reason. Purely expressive action may sometimes count as intentional, but the passage to it from emotion would not seem to make sense as an implicit inference. An expressive act might sometimes be seen as instrumental to some general end, such as discharging bottled-up feelings or communicating one's feelings to others, but that would not explain all cases. Crying need not alleviate sadness, or jumping compound joy, and sometimes one wants to keep such reactions private.

I spent little time on positive emotions, as rewarding action with pleasurable affect. Action from joy or gratitude may serve to sustain the feeling, or to reinforce it with further positive feelings, but I thought this too obvious to require argument. What seems more important to note was that emotions that evaluate their objects positively may have negative elements, typically of unsatisfied desire. Love, say (or what I call "attachment-love"), in a case of distance from the love-object, may involve discomfort at the thought that one ought to move closer. Gratitude may involve uneasiness at

one's inability (so far) to repay the object of gratitude. Any element of *felt need* in an emotion, even a need for expression, can bring it under my account in terms of emotional discomfort. Only an emotion that exerts no pressure to act, on the order of tranquil joy, would serve simply as an end. But then one would not act *from* it but just in order to attain or sustain it. The occurrent emotion would not itself provide a practical reason. I now want to suggest a way of looking at reasons that yields a somewhat broader view.

4 Discounting reasons

I began work on reasons with no thought of linking my argument to emotions, and I think the view I came up with is best explained independently. My main aim was to question the link some authors assume between reasons and rationality, such that failure to act on an acknowledged reason that is not opposed by another at least as strong – an all-things-considered reason that the agent is aware of as such – would be irrational (see, e.g., Williams 1981). But discounting acknowledged reasons – simply setting them aside in practical reasoning, without appeal to contrary reasons – need not be irrational. Other authors have defended similar claims (e.g., Gert 2003; Dancy 2004), but I sought a deeper explanation in the nature of reasons generally that would do fuller justice to rational discounting, in particular by allowing for discounting some practical requirements.

I call the view I came up with the "critical" conception of reasons, since it takes reasons *against* action – reasons that lodge some criticism of action – as primary, with reasons that instead simply count in favor of action understood in the first instance as answering potential criticism (Greenspan 2007). It is reasons lodging criticism that ground requirements to act, by ruling out alternatives to action, much as moral requirements rest on prohibition of alternatives. By contrast, standard conceptions of reasons focus on reasons in favor of action (e.g., Scanlon 1998, ch. 2) so that reasons against action are seen, in effect, as reasons in favor of omitting it. I think this focus may be a result of the origin of reasons-talk in talk of motives, but, among other things, it encourages the notion that unopposed or strong enough reasons add up to a rational requirement. That is what I question.

In past work, I often refer to reasons in favor and reasons against as positive and negative reasons, but since many of the reasons I think of as negative (particularly those underlying requirements) tend to be stated in positive form, I eventually dubbed them "critical" reasons. My reason for eating a healthy diet, say, implicitly lodges a criticism of too much junk food as bad for health and hence amounts to a critical reason – whereas, if chicken is healthy, that is a point in favor of eating it, but assuming that a healthy diet need not include chicken, my reason to eat it is merely a "favoring" reason, what I originally called a "purely positive" reason.

In this chapter, I also want to avoid confusion with the positive or negative valence of emotions. A practical reason can also be seen as evaluating something, but the "something" in question has to be an action (or inaction), either recommended or ruled out. Some emotional evaluations, such as the unfulfilled ought-judgments I associate

with emotional desire, also concern action, so there will be important overlap with reasons. But it would be best to avoid terminology that suggests overlap in all cases.

What discounting a reason amounts to depends on which sort of reason it is, favoring or critical. In general, I take "discounting" a reason to mean assigning it no weight in one's practical reasoning. The question is whether one needs to appeal to a further reason justifying discounting in order to do so legitimately. A reason that lodges no criticism at all of alternatives to action simply recommends an option and thus can be discounted at will. It gives a point in favor of some action and in that sense amounts to a favoring reason, but it does not favor that action *over* others. Consider my reason for having a snack right now – in the absence of a craving or other felt need, but just because I think I would enjoy one. Assuming that this is the only reason relevant to my choice of action – there is no objection to either my having or my forgoing a snack right now – I can simply decline to take it into account. It still counts as a reason, and I still consider it one; indeed, it would seem to be an all-things-considered reason. But it would not be irrational for me to ignore it. It essentially offers me an opportunity I can turn down.

On the other hand, I could not so easily discount a critical reason for action. Suppose that my health requires forgoing large meals at the usual times and instead snacking on healthy foods at smaller intervals throughout the day. Essentially, this means there are objections to my waiting too long to take in nourishment – and at this point, having postponed a snack for some time, there is a critical reason against postponing one further. Or consider the reason I have for exercising regularly, which amounts to or includes a reason against prolonged inactivity. Now, either reason can be overridden by more important considerations: perhaps I have social obligations involving a meal and would do better not to snack beforehand, or perhaps a doctor has banned vigorous exercise on a day after some minor surgery. But discounting means ignoring a practical reason, setting it aside, on the basis of a decision to do so, not because it loses out to opposing reasons.

If all practical alternatives but one are subject to criticism, rationality would seem to require either taking the remaining option or answering the criticism with a favoring reason that provides justification for turning that option down. However, it sometimes seems reasonable to make and act on a decision to stress certain reasons over others, not on the basis of their pre-given weights – some independent notion of their comparative importance – but as a matter of *setting priorities*. Consider the reasons I might have for getting out of the house on a particularly nice summer day. The weather is perfect, so I certainly have a favoring reason. If we suppose that I have been spending an unusual amount of time indoors, working on the first draft of a paper, I also have a critical reason, a reason against missing a rare spell of not-too-humid summer weather. All things considered, no doubt I ought to get out, but I have resolved not to interrupt my daily writing schedule until I have finished the draft. My decision essentially gives me a higher order reason to discount reasons against adhering to my schedule, short of emergencies. This higher order reason need not be deemed *more important* than

my reasons for getting out – I am not facing a deadline or the like – but it takes precedence over them by virtue of my decision. My decision may or may not be *ideally* rational, or rational in the sense of maximizing my advantage, but I am within my rights, rationally speaking – it is rationally permissible for me – to take charge of my reasons in the way I have done, essentially waiving my criticism of a certain option.

Of course, there are limits. I must get out at some point, possibly in less inviting weather. Again, let us bypass matters of degree. But note that I do still have a reason to get out specifically today, despite my decision to discount it. Discounting it means declining to consider it in practical reasoning about what to do today, not rescinding its status as a reason. I also assume that rationally permissible discounting is limited to self-regarding critical reasons: I lack the authority to waive the criticisms of my actions by others that ground typical moral requirements, assuming that I recognize them as such (see Greenspan 2007). But in contrast to favoring reasons, which involve no criticism, my decision to discount a critical reason requires justification, if only by the need to set priorities. Perhaps I think that an attempt at balance in my summer activities would interfere with my momentum in writing the draft, for which I need a solid block of time. Or perhaps it is not a question of what I need but just of what I prefer, so that my higher order reason is itself just a favoring reason. In that case my higher order reason might be seen as making my critical reason optional – at least until I make a definite decision, which then yields a critical reason against failing to follow through.

In short, then, practical reasoning sometimes includes a further level, of reasoning *about* one's reasons, that can overturn the results of a first-order comparison of reasons by weight. In deciding which of two competing reasons to discount (or at least to postpone acting on – to discount in application to my present action, that is), an agent might be said to be weighting (with a "t"), rather than merely weighing, her reasons. She is putting her thumb on the scales, as it were, in response to a higher order reason, but possibly one whose optional status gives her the final say. If I am right in thinking this can be rational, then rationality includes not just responding to relevant reasons but also actively shaping them in appropriate ways.

5 Emotions and higher order reasons

A rational choice, in the sense of one that is "within reason," or rationally permissible, may or may not be wise, or ideally rational. Perhaps I have reason to think that staying in for the stretch of time I need to finish a draft will have some longer-term ill effects – on my mood, when the new term begins, say, and hence on the likely success of my courses. So I am not just discounting the expectation of pleasure outdoors, if I discount that reason – though it might seem less important to me now than getting a draft done over the summer, when I have an adequate block of time. I think we may grant that, whichever choice I make, I would not be irrational. This is, as they say, a "judgment call."

Here is where emotions can play a crucial role in practical reasoning, as supplying further reasons against discounting – normative, not just motivating, reasons, insofar as they involve good or bad affective states of the agent, states worth sustaining or alleviating. On the view I sketched earlier, emotions can serve to register practical reasons in affect, thereby providing further reasons to change situations (or avoid acts) evaluated negatively. Suppose that, besides simply recognizing the possible consequences of not getting out much this summer, I also am a bit worried about them. As a variant of fear, worry is an unpleasant and distracting state that threatens to continue as long as I discount my other reasons for getting out. By bringing home to me in present terms something I might otherwise dismiss as a problem to be dealt with later, it serves as a barrier to discounting,

Now, one might think that in the circumstances described I would have sufficient reason against discounting without any special role for my current state of feeling. If nothing else, what I am likely to feel later, when my courses do go poorly, should weigh no less with me than what I feel now – not to say the fact that my courses do go poorly. However, though I have reason now to take account of my later reasons, the question is what I am justified in believing as things now stand. Perhaps it also seems possible that a summer spent largely indoors, plugging away at my writing, will make me particularly eager to interact with students once the term begins, thus enlivening my teaching. Still, it would be unwise to count on that. So worrying enough to keep in mind the possible negative consequences of simply discounting my reasons against staying in may still play a useful role in practical reasoning. As I have argued in general terms, an emotion can be justified as a way of focusing attention on a generally

significant evaluation, even in the absence of adequate evidence for the corresponding judgment.

My argument that emotions supply reasons for action – or in this case, for an omission: declining to discount certain reasons – assumes a negative affective element of emotion as providing a reason to falsify the emotion's evaluative content. So how might a purely positive emotion like joy figure in, except as a goal of action? Remember that I do not deny that nonemotional awareness of a reason can be enough to prompt action. So if an agent currently experiencing joy is reluctant to lose the feeling, her joy can ground a critical reason, against changing the situation that prompts the emotion, even if her awareness of that reason involves no element of negative affect. Suppose that, after a month indoors slowly writing a draft, I finally fix on the answer to a problem I needed to solve. Of course, the thought that what I write in this state will be good may give me reason to stay in and keep writing. But the fact that I *feel* so good also supplies a reason in itself, insofar as it makes me critical of alternatives that would fail to sustain the feeling. There are important limits here, as usual, but the feeling at least serves to reinforce my other reasons against interrupting my writing.

In short, then, to the extent that a positive occurrent emotion can supply a critical reason – a reason against failing to sustain it – it does have a role to play in practical reasoning. But the main role of emotion in my account is still reserved for negative affect. Given that emotional discomfort threatens to continue until one acts to change the situation, it can supply a higher order reason against discounting reasons for timely action on an ought-judgment that requires action in general terms but would otherwise allow for postponement – or, conceivably, for other options that might make action less effective. But further, if an emotion can sometimes be rationally justified as a way of holding in mind a reason of general importance in cases where one lacks adequate evidence for the corresponding judgment, and emotional discomfort itself provides a reason for action to relieve it, as I have argued, then negative emotion can sometimes do more than reinforce independent judgments about one's reasons. Sound practical reasoning can sometimes issue from appropriate emotions in the absence of justified belief in the corresponding judgment.

To see this, consider again the case where I am worried about the long-term effects of staying indoors too long this summer. Suppose I do not really have (or think I have) adequate grounds for believing such effects are all that likely – I have a month more to enjoy the outdoors between finishing a draft and starting the new term – but the mere possibility seems enough to justify some concern. Maybe the weather will be so hot then that I will want to stay in. Assuming my worry is not disproportionate, it is justified as a way of holding in mind a negative evaluation of staying indoors any longer (my critical reason) in the absence of a justified judgment to that effect. But its element of discomfort itself gives me some reason to get out – if only to alleviate my worry.

At this point, one might well ask: why not discount the emotion, or any reasons it gives rise to? I might just tell myself to ignore my worry about long-term effects and

push on with my writing schedule until I get a draft done. This is indeed an option but in the present case no more than that. If the emotion posed practical problems – kept me from concentrating on my writing, say – I might be *required* to discount it, but not just because I lack sufficient evidence for the corresponding evaluative judgment. However, in the case as described, though the effects of remaining indoors are far from certain, some worry about them is warranted, and my present level of worry is assumed to be unproblematic. Warrant, or justification, is a threshold notion: having enough need not yield a requirement but just an option. In fact, my initial purpose in proposing the critical conception of reasons was to allow for optional reasons – even reasons of the sort that normally yield requirements but can be *rendered* optional by the agent's choice to discount them. Emotional discomfort adds a further reason against failing to act, but it does not compel action. Its potential for distraction just makes it harder to ignore than a non-emotional reason. Setting it aside requires a bit of effort, so it exerts a degree of pressure to act, and to act soon.

5 Conclusion

Let me pull things together very briefly. On the account of emotions and reasons I have proposed, emotion can play a valuable role *within* practical reasoning, not just as an alternative to it, insofar as emotional affect serves both to register and to reinforce the act-evaluations that constitute practical reasons. That is to say that it both holds them in mind and rewards or punishes action or failure to act on them. Sometimes it can do so, and be justified in doing so, in the absence of adequate warrant for the corresponding practical judgment. Moreover, emotional discomfort – and sometimes even unalloyed positive emotion – can give rise to critical reasons, reasons against practical alternatives, of the sort that yield requirements. Emotion can therefore supply a barrier to discounting reasons that could otherwise legitimately be set aside.

This is not to say that we deliberate about our emotions, at least in normal contexts. Rather, they serve as background factors in practical reasoning: good or bad states of an agent that guide her assessment of her reasons for action without her having to reflect on them. They direct attention elsewhere by virtue of their evaluative content. Though their justification as appropriate depends on warrant for their evaluative content, I identify them, not with evaluative thoughts, but with the states of affect that do the evaluating and thereby augment our reasons.

References

Aristotle (1941) "Rhetoric," 1325–1451, in R. McKeon (ed.), *The Basic Works of Aristotle*, New York: Random House.

Arpaly, Nomy (2003) *Unprincipled Virtue: An Inquiry into Moral Agency*, Oxford: Oxford University Press.

Dancy, Jonathan (2004) "Enticing Reasons," in R. J. Wallace, P. Pettit, S. Scheffler, and M. Smith (eds.), *Reason and Value: Themes from the Moral Philosophy of Joseph Raz*, Oxford: Oxford University Press.

Gert. Joshua (2003) "A Functional Role Analysis of Reasons," *Philosophical Studies*: 1–26.

Greenspan, Patricia S. (1980) "A Case of Mixed Feelings: Ambivalence and the Logic of Emotion," in A. O. Rorty (ed.), *Explaining Emotions*, Berkeley: University of California Press.

———. *(1988)* Emotions and Reasons: An Inquiry into Emotional Justification*, New York: Routledge, Chapman and Hall.*

———. (2003) "Emotions, Rationality, and Mind/Body," *Philosophy* 52: 113–25.

———. (2007) "Practical Reasons and Moral 'Ought'," in Russ Schafer-Landau (ed.), *Oxford Studies in Metaethics*, Vol. II, 172–94, Oxford: Oxford University Press.

———. (2010) "Learning Emotions and Ethics," in P. Goldie (ed.), *The Oxford Handbook of Philosophy of Emotion*, Oxford: Oxford University Press.

———. (2012) "Craving the Right: Emotions and Moral Reasons," in C. Bagnoli (ed.), *Morality and the Emotions*, Oxford: Oxford University Press.

———. (2019) "The Evaluative Content of Emotion," *Philosophy* 85: 75–86.

Haidt, Jonathan (2001) "The Emotional Dog and Its Rational Tail," *Psychological Review* 108: 814–34.

Hume, David (1978) *A Treatise of Human Nature*, Oxford: Clarendon Press.

Roeber, Blake (2016) "The Pragmatic Encroachment Debate," *Nous*: 1–25.

Scanlon, T. M. (1998) *What We Owe to Each Other*, Cambridge, MA: Harvard University Press.

Solomon, Robert C. (1993) *The Passions: Emotions and the Meaning of Life*, Indianapolis: Hackett.

Sousa, Ronald de (1987) *The Rationality of Emotion*, Cambridge, MA: MIT Press.

Stanley, Jason (2005) *Knowledge and Practical Interests*, Oxford: Oxford University Press.

Williams, Bernard (1981) "Internal and External Reasons," in *Moral Luck: Philosophical Papers 1973–1980*, Cambridge: Cambridge University Press.

18. Psychopathy, agency, and practical reason

Monique Wonderly

Psychopaths exhibit a philosophically interesting combination of deficits and rational competences. They appear perfectly capable of basic means-end reasoning. They typically perform at or above average on intelligence tests, and many are quite adept at employing sophisticated forms of reasoning in order to skillfully manipulate others. Still, psychopaths seem profoundly impaired in at least one important domain of practical reasoning, moral reasoning. Many theorists have argued that psychopaths lack the capacity to competently recognize and/or respond to moral reasons.

Theorists specify and explain this impairment in a variety of ways. Some afford explanatory roles to certain emotional abnormalities, such as difficulties empathizing or experiencing negative affects. Others point to deficits that are not necessarily construed as emotive, such as an inability to recognize the interests and/or authority of others as intrinsically reason-giving, an inability to value, insufficiently unified agency, lack of self-control, or general learning deficiencies. As theorists have noted, many of these features are not merely relevant to moral reasoning, but they also bear on the capacity for practical reasoning more broadly.

Philosophers have urged that considerations about the psychopath's capacity for practical rationality can help to advance some metaethical debates. These debates include the role of rational faculties in moral judgment and action, the relationship between moral judgment and moral motivation, and the capacities required for morally responsible agency. In what follows, I discuss how the psychopath's capacity for practical reason features in these debates, and I identify several takeaway lessons from the relevant literature. Specifically, I show how the insights contained therein can illuminate the complex structure of practical rationality, inform our standards for an adequate theory of practical reason, and frame our thinking about the significance of rational capacities in moral theory and social practice.

1 Psychopathy

Let's first consider what the term *psychopath* is meant to capture.[1] We are not, for example, interested in the "psychopath" that we often find in popular cinema: the blood-thirsty sadist, bent on doing evil for evil's sake. Actual psychopaths are rarely so committed and not especially sadistic. However, they typically do engage in dangerous, and sometimes violent, behavior.

Psychopathy is a personality disorder that involves a tendency toward antisocial behaviors (e.g., threatening behavior such as verbal abuse or violence, repeated criminal conduct) and certain emotional-interpersonal deficits (e.g., shallow affect, lack of empathy, inability to feel guilt). The most commonly used diagnostic tool for identifying psychopathy is a revised version of the Psychopathy Checklist (PCL-R). The PCL-R scores subjects according to the extent that their behaviors and attitudes reflect the twenty items on the checklist. The items – characteristics such as superficial charm, callousness, impulsivity, grandiose sense of self-worth, manipulative behavior, lack of realistic long-term goals, delinquency, and so on – are measured on a 0–2 point scale, and a score of 30 suffices for a psychopathy diagnosis. The average person scores about 4. Psychopaths make up less than one percent of the total population, and the vast majority are male (Kiehl and Buckholtz 2010).[2]

Psychopaths do not usually suffer from compulsions, delusions, or any obvious intellectual impairments. They can be brilliant and charismatic. Some are successful at evading criminal prosecution and occasionally achieve positions of wealth and power. In some respects, psychopaths blend in with society rather well. But they diverge in others. They are quick to anger and often lash out fiercely when others interfere with their aims. They are also known to coolly and remorselessly employ violence as a means to achieving their goals (Blair 2008).

Though psychopaths are sometimes superficially charming, they are often considered incapable of (and uninterested in) meaningful relationships. According to some clinicians, they lack the capacity to love (Cleckley 1976; Hare 1993). Despite their

[1] There is a worry that in making general – and, in particular, *moral* – claims about "the psychopath," we risk inappropriately marginalizing large groups of actual people. That is not my intention. I take it that not all who have been diagnosed with psychopathy neatly fit the following criteria, and so we should be cautious about indiscriminately extending potentially harmful judgments based on this model to those who identify as psychopaths. Also, it is not without some regret that I use the term "psychopath" here, as opposed to "psychopathic individual," and I do so only for the purpose of maintaining continuity and cohesion with the relevant literature. I thank Hanna Pickard for helpful discussion on this point.

[2] Consequently, I will use masculine pronouns when referring to psychopaths.

egocentricity and intelligence, they often engage in impulsive behavior that is not only harmful to others but antithetical to their own interests. Robert Hare offers an example of such behavior when he describes the actions of one psychopath who decided to stop for beer on his way to a party but realized he'd forgotten his wallet. Hare explains, "Not wanting to walk back, he picked up a heavy piece of wood and robbed the nearest gas station, seriously injuring the attendant" (1993, 59).

While psychopaths perform comparatively well on most cognitive measures – including IQ tests and (some) moral reasoning tasks, they have tended to do poorly on one task that four year-old children tend to perform with relative ease, Elliot Turiel's Moral/Conventional Distinction Task (MCT). The MCT measures one's ability to distinguish between moral and conventional norms, where transgressions of the former are taken to be more serious, less permissible, and less authority dependent (Turiel 1983). Examples of typical moral transgressions include hair pulling and hitting, while examples of conventional transgressions might include drinking soup from the bowl and wearing pajamas to school. Although recent research indicates that psychopaths perform relatively well on a modified version of the MCT, earlier studies suggest that psychopaths have considerable difficulties identifying the relevant distinction (Blair 1995).[3]

Certain conversational oddities seem to reflect peculiar gaps in the psychopath's understanding. Consider one psychopath who, having just described how he kidnapped a woman, repeatedly raped her, and slit her throat, said to his interviewer: "Do you have a girl? Because I think it's really important to practice the three C's: caring, communication and compassion. That's the secret to a good relationship. I try to practice the three C's in all my relationships" (Kiehl and Buckholtz 2010, 25). Another psychopath referred to his mother as the "most beautiful person in the world," confessed to stealing her jewelry as a child, and then added, "You know, I never really knew the bitch – we went our separate ways" (Hare 1993, 40).

While psychopaths often use moral and emotional terms, their conversational exchanges often suggest that they don't really understand the concerns and emotional experiences of others. One psychopath boasted about how he benefitted his rape victims: "There'd be interviews with the victims. They'd get their names in the paper. Women, for example, would say nice things about me, that I was really polite and considerate, very meticulous ... Some of them thanked me" (Hare 1993, 43). Or again, consider the words of another rapist, high on the PCL-R, who admitted confusion over his victim's feelings: "They are frightened, right? But, you see, I don't really understand it. I've been scared myself, and it wasn't unpleasant" (ibid, 44).

These apparent deficits in understanding distinguish psychopaths from average violent criminals. To this, we can add that psychopaths also frequently exhibit distinctive

[3] For a discussion of research indicating that psychopaths perform well on a modified version of the MCT, see Aharoni et al. 2012. Some theorists remain unconvinced that the newer studies conclusively demonstrate the psychopaths' (unimpaired) facility with the relevant distinction, since the amended version of the task seems notably easier than the original version (see, for example, Kumar 2016).

physiological abnormalities. They show reduced responsiveness to emotional stimuli and often fail to exhibit the normal bodily reactions associated with registering the need to reverse one's behavior in light of negative outcomes (Blair 2008). Some imaging studies suggest that psychopaths have underdeveloped neural structures in brain areas involved in processing emotions, motivation, and self-control (Kiehl and Buckholtz 2010, 27).

With this sketch in hand, we can now turn to the role of the psychopath's capacity for practical reason in philosophical discourse.

2 Psychopathy and practical reason in philosophical discourse

As I indicated in the introduction, considerations about the psychopath's capacity for practical reason have featured in at least three domains of philosophical investigation: (1) the role of rational faculties in moral judgment and action, (2) the relationship between moral judgment and moral motivation, and (3) the capacities required for morally responsible agency. The first area concerns the dispute between moral rationalists and moral sentimentalists. The second concerns the debate between motivational judgment internalists and motivational judgment externalists. And the third concerns two competing views about the capacities required for moral responsibility: the rational competence view and the moral competence view.

2.1 Moral rationalism and moral sentimentalism

The debate between moral rationalists and moral sentimentalists is a nebulous one. Historically, it has concerned the order of priority, and respective roles, of reason and emotion in morality. As some have framed the guiding question: Is morality more like math or beauty (Gill 2007)?

Moral rationalism is associated with the claim that morality is a species of practical rationality. As Jeanette Kennett helpfully puts the point, moral rationalists are minimally committed to the claims that certain rational capacities are essential to moral agency and moral judgments are judgments about reasons – that is, in judging it wrong to Φ, an agent takes herself to have identified a consideration that counts against Φ-ing (2010, 251).[1]

Moral sentimentalists, on the other hand, typically afford pride of place to sentiment and emotional capacities in moral judgment and action. Some argue that empirical facts about psychopathy support a sentimentalist, rather than a rationalist, view of morality. The psychopath appears to be morally impaired, despite having (what are often taken to be) intact rational abilities. Given psychopaths' well-documented emotional deficits, sentimentalists can easily explain this, but not so for moral rationalists. Or so the challenge goes.

[1] See also Michael Smith 1994.

Jesse Prinz (2006), for example, advances a view on which emotions are both necessary and sufficient for moral judgments, adducing findings about psychopathy to support the necessity claim. He characterizes psychopathy as "the perfect test case" for the thesis that emotions are necessary for moral development. Prinz argues that owing to deficits in core negative emotions, such as fear and sadness, psychopaths cannot acquire empathetic distress or guilt and are consequently unable to make moral judgments (2006, 32). In his view, the relevant emotional deficits constitute the "root cause" of their antisocial behavior (ibid).

Shaun Nichols (2002, 2004) takes a similar approach. According to Nichols, the empirical evidence suggests that psychopaths have impaired capacities for making moral judgment, but contra what rationalists might have us expect, the relevant deficit is a defective emotional response. Nichols cites psychopaths' difficulty with the Moral/ Conventional Distinction Task as evidence of moral impairment and the clinical description of the psychopath's emotional deficits as evidence of emotional impairment. He argues that psychopaths' emotional deficits inhibit their ability to make "core moral judgments," where core moral judgments "are judgments that implicate 'Sentimental Rules,' or rules prohibiting actions that are independently likely to elicit strong negative affect."[2] In his account, rationalist views seem ill suited to explain the psychopath's poor performance on the MCT, given that psychopaths seem rationally competent and very young children and many individuals with severe cognitive disabilities do well on the task (2002, 296).

Moral rationalists have sought to defuse the sentimentalist challenge by arguing that: (1) evidence that emotions play important roles in moral judgment and action does not undermine (all varieties of) moral rationalism, and (2) psychopaths have deficits in practical reason that might explain their moral impairments.

In support of (1), Kennett argues that rationalists can comfortably accommodate a significant role for sentiments in morality, noting that Immanuel Kant regarded "moral feeling" as a subjective precondition for "receptiveness to reason's constraints" (2002, 354). Kennett suggests several rationalist-friendly interpretations of the psychopath's emotional-moral impairment. Psychopaths seem to lack reverence for reason or, again, displeasure at cognitive dissonance – affective phenomena that may facilitate susceptibility to, and effective employment of, reason (2002, 355, 2010, 256).[3] In a similar vein, Heidi Maibom suggests that emotional deficits might impact the psychopath's moral abilities "through practical reason alone," thus leaving moral rationalism unscathed. If, for example, psychopaths' deficits in negative emotions inhibit moral education

[2] Nichols explains that core moral judgments are guided by an "internally represented body of information, a 'normative theory' prohibiting behavior that harms others" (2002, 16) and "some affective mechanism that is activated by suffering in others" (ibid, 18).

[3] Kennett, following her interpretation of Kant, suggests that reverence for reason, understood as "the concern to act in accordance reason," is the core moral motive and suggests that the psychopath's "indifference to reason is the key to his behavior" (2002, 355). For an insightful response to Kennett on this point, see Victoria McGeer 2008.

"through reward-punishment centered learning," as Prinz suggests, then a rationalist interpretation of the relevant moral impairment seems apt, notwithstanding the role played by emotion (2010, 8). After all, the ability to acquire knowledge via reward-punishment–centered learning is a rational ability.

Moral rationalists have also sought to defend their position by arguing directly for (2), the view that psychopaths also have *rational deficits* that might explain their moral impairment. Maibom, for example, highlights a host of cognitive difficulties that might inhibit the psychopath's capacity for practical reason. Psychopaths are notoriously impulsive, have inflated notions of their abilities, and have trouble adjusting their behavior in light of negative outcomes. In Maibom's account, these attentional problems, inaccurate self-estimates, and "reversal deficits" negatively impact the psychopath's abilities to will the necessary and sufficient means to his ends, to coordinate his intentions with foreseeable consequences of his actions, and to coherently universalize maxims (2005, 239, 247). Maibom concludes that as these factors bear on one's competence in grasping and enacting moral duties, the psychopath's practical irrationality plausibly explains his moral impairment.

Similarly, Kennett argues that the psychopath is at best short-term instrumentally rational, lacking any coherent, extended conception of his ends and often failing to adopt reasonable means toward achieving his proclaimed ends (2002, 2006). The psychopath might profess to have lofty goals, but his goals are often both fleeting and unrealistic. Within the same day, a psychopath might form and give up his plan to become a professional athlete, deciding now to become a doctor – all while lacking the requisite skills and training (and any plans to acquire them) for either. Kennett also points out that psychopaths often employ "grossly disproportionate means to their immediate ends" (2006, 76). Think here of the psychopath who, wanting to buy beer for a party, decided to brutally assault and rob a service-station attendant rather than return home for his forgotten wallet.[4]

Marko Jurjako and Luca Malatesti (2016) review experimental results concerning the psychopath's performance on instrumental learning tasks, and they argue that peculiarities in the psychopath's performance are likely attributable to informational inaccessibility rather than impaired instrumental rationality. In other words, as they interpret the evidence, psychopaths' insensitivity to certain kinds of information may, in some cases, render them unaware of the means for their ends but not incapable of "willing the accessible means that are necessary or sufficient for accomplishing some end" (2016, 726).

Doubtless, psychopaths often successfully engage in means-end reasoning. Importantly, though, Kennett suggests that basic means-end reasoning does not suffice for rational agency, the "markers" of which include the capacities for normative reflection and rational self-control (2010, 254). She writes,

[4] Kennett and Steve Matthews discuss this case in a later work, describing the psychopath's actions as "not just immoral" but "stupid" (2008, 225).

> A person who could not reflect upon whether or not his desires provided 'reasons' for action, whose desires were entirely unresponsive to such reflection, or who could not be guided by the results of his deliberations, through exercises of planning and self-control, would not count as a rational agent. (ibid, 252)[5]

The psychopath, being severely limited in his abilities to take an evaluative perspective on his desires and to delay their immediate satisfaction, seems to exemplify this rational defect. Drawing on child development research, Kennett explains that the capacities for normative reflection and rational self-control typically begin to emerge around the same time as the capacity for making the moral/conventional distinction, thus suggesting a link between psychopaths' rational deficits and their moral impairment (ibid, 253).

2.2 Motivational judgment internalism and motivational judgment externalism

Considerations about psychopathy have also entered into the debate between motivational judgment internalists and motivational judgment externalists. In its barest form, motivational judgment internalism (MJI) posits a necessary connection between moral judgments and moral motivation (Smith 1994; Roskies 2003). MJI is often construed in terms of the following thesis: If Agent A judges that Φ-ing is morally wrong, then necessarily A will be motivated not to Φ, or, again, if A judges that s/he morally ought to Φ, then necessarily she will be motivated to Φ.[6]

Motivational judgment externalism (MJE) is just the denial of MJI. According to motivational judgment externalists, moral judgments do not entail corresponding motivations. A common objection against MJI is the possibility of a "rational amoralist," someone who knows and understands the dictates of morality (and presumably makes moral judgments) but doesn't care about morality and is unmotivated to comply with moral norms (Brink 1989). The psychopath – who, in some understandings, seems similar to the rational amoralist – might then pose a challenge for proponents of MJI.[7]

[5] Importantly, not all would agree that the defects Kennett identifies here constitute practical irrationality. First, one might draw a distinction between structural rationality and responsiveness to reasons (see, for example, John Broome's entry in this volume). Second, even those who associate practical rationality with reasons-responsiveness might deny that failure to engage in normative reflection represents irrationality, as opposed to a mere failure to exercise one kind of rational capacity. Thanks to Kurt Sylvan for prompting me to highlight this point.

[6] Some theorists have offered weaker formulations. Michael Smith's preferred version holds that if Agent A judges that Φ-ing is wrong, then either A will be motivated not to Φ or A is practically irrational (1994, 61, 2008, 211).

[7] Adina Roskies (2003) cited research findings on a group of patients with "acquired sociopathy"

Internalists have typically responded by insisting that psychopaths (sociopaths, etc.) don't *really* make moral judgments but only do so in "an inverted commas sense" and thus pose no threat to MJI (Smith 1994).[8]

As internalist replies to the psychopathy objection tend to employ many of the same arguments canvassed in section 2.1, we can afford to be relatively brief here. In support of their claim that psychopaths do not make moral judgments, theorists often cite the psychopath's poor performance on certain empirical measures and oddities in his use of moral language. Kennett and Cordelia Fine (2008), for example, discuss myriad studies that, in their account, suggest that psychopaths do not make genuine moral judgments. Familiarly, they note that psychopaths have significant difficulties drawing the moral/conventional distinction, forming value judgments without making exceptions of themselves, and correctly deploying moral concepts in conversation (2008, 174–178).[9] Kennett and Steve Matthews (2008) raise similar points, adding that the psychopath's more general rational impairment renders him a poor candidate for the "rational amoralist." As they argue, psychopaths are only "very implausibly viewed as rationally unified agents," given their lack of facility with normative reasons (2008, 222, 224). Owing to poor self-regulation skills, shortened attention spans, and impoverished conceptions of their own well-being, psychopaths are unable to grasp, and to guide themselves by, the normative considerations that typically unify and sustain extended agency. We find evidence for this in their self-destructive behaviors and erratic, contradictory speech (ibid, 223–224). Kennett and Matthews urge that these deficits are likely related to psychopaths' moral deficits, since moral agency and rational agency apparently go together in the normal case. Adducing child development studies, they conclude that since the higher-order cognitive capacities required for self-constitution and extended agency are the "same capacities that make us rationally susceptible to moral claims," it seems unlikely there exist rational amoral agents (ibid, 228).

Some have expressed skepticism about the psychopath's supposed inability to make genuine moral judgments. In responding to Kennett and Fine, for example, Roskies (2008) argues that the evidence suggests that psychopaths reason *differently* about moral norms but not that they altogether fail to make moral judgments. Roskies also doubts that the ability to draw moral/conventional distinctions is necessary for making moral judgments. She writes, "Psychopaths are still cognizant of what is morally right

---

as evidence against MJI. According to Roskies, following injuries to the ventromedial prefrontal cortex area of the brain, these patients continued to have normal moral beliefs and to make moral judgments, but they were no longer *inclined to act in accordance with* those beliefs and judgments, thus falsifying MJI (2003, 63).

[8] We find sentimentalists on both sides of the debate. Prinz, for example, suggests that psychopaths "furnish internalists with a useful piece of supporting evidence," insofar as their co-occurrence in moral motivation and moral competences appears to be linked (2007, 44). Nichols argues that considerations about psychopathy suggest against some varieties of internalism, including "conceptual judgment internalism about moral judgment" and "empirical internalism about harm-norm judgment" (2004, 109–115).

[9] In a later work, Kennett argues that psychopaths lack competence with moral concepts, as they are not "conversable" with the relevant terms (2010, 246).

and wrong... . Even if their concepts are impaired, it is plausible that they are nonetheless moral concepts" (2008, 202). Relatedly, Walter Sinnott-Armstrong (2014) argues that the empirical evidence concerning psychopaths' abilities to make moral judgments is inconclusive. He suggests that the relevant studies often suffer from methodological limitations and sometimes yield mixed results – noting, for example, that psychopaths tend to perform well on a modified version of the MCT (2014, 195).[10]

Much of the dialectic turns on whether internalists and externalists can agree upon criteria for determining whether a moral judgment has been made that doesn't *presuppose* the presence or absence of moral motivation. Some have proposed that we can make headway by acknowledging that facility with moral concepts is at least necessary for moral judgment or, again, that we can use the MCT to help identify the key features of a distinctly moral judgment.[11] As we have seen, however, theorists have questioned both the criteria themselves and the validity of the tools used to measure them.

2.3 Morally responsible agency: rational competence versus moral competence

Theorists have employed considerations about psychopathy to illuminate another important metaethical issue: the matter of which capacities are required for morally responsible agency.

Though many creatures might harm us, we only hold some of them morally responsible for their harmful actions. For example, lions, toddlers, and those in the grips of psychotic delusions may sometimes inflict harm, but we typically do not *blame* them for doing so. They are unable to fully grasp the nature of their actions and to exercise rational control over their behaviors. Thus, they are not appropriate candidates for the blaming and praising attitudes (e.g., resentment and gratitude) by which we hold one another morally responsible.

Psychopaths represent a more perplexing case. At the very least, psychopaths have some understanding of societal norms and how to comply with them. They know that theft and assault are grounds for legal punishment and sometimes show self-restraint in order to avoid such penalties. What's more, psychopaths often demonstrate awareness of moral expectations. They know, for example, that helping others is generally considered morally good while deceit is considered wrong. They sometimes use this knowledge to manipulate others, representing themselves in conversation as "generous" or "honest" to get what they want. These competences seem to set psychopaths apart from lions, toddlers, and those suffering from psychotic delusions.

[10] For discussion of recent empirical work on psychopathy and moral judgment, see Sinnott-Armstrong 2014, 193–199.

[11] See, for example, Kennett 2010; Kumar 2016.

In some accounts, psychopaths' rational capacities render them eligible candidates for blame. T.M. Scanlon, for example, claims that "a rational creature who fails to see the force of moral reasons" might be properly subject to moral criticism, provided he can "understand that a given action will injure others and can judge that this constitutes no reason against so acting" (1998, 288). Similarly, Matthew Talbert argues that despite being "morally blind," psychopaths are blameworthy for their actions because they are "capable of making decisions on the basis of judgments about reasons" (2008, 519). Talbert describes psychopaths as effective practical reasoners who can "count the pleasure of having a possession of yours as a reason to take it from you and ... form the judgment that nothing about the effect of this action on you is a reason to refrain from performing it" (2008, 522). In these views, owing to psychopaths' rational competence, their actions can express offensive judgments that legitimize blaming attitudes.[12]

Many theorists reject the view that *mere* rational competence – that is, facility with reasons in general – suffices for moral accountability. Gary Watson, for example, is among those who argue that moral accountability requires competence with *moral reasons* in particular. In this view, an otherwise rationally competent psychopath would be exempt from moral responsibility if he were unable to recognize moral reasons.[13] This claim raises questions about the particular capacities required for the kind of recognition at issue. What must morally competent agents be able "to do" with moral reasons? And why should we think that psychopaths lack the relevant abilities? Watson employs an argument from moral communication (henceforth, AMC) to answer these questions.

Watson argues that resentment, along with other reactive attitudes by which we hold others morally accountable, "are incipiently forms of communication, which make sense only on the assumption that the other can comprehend the message" (1987, 264). He explains that resentment expresses a moral demand for reasonable regard and suggests that very young children and psychopaths may lack sufficient moral understanding to be proper recipients of the relevant demand (ibid, 271).[14] Watson later elaborates on this position, describing psychopaths as "unreachable by the language of moral

[12] In a recent modification and extension of her earlier (2003) work, Patricia Greenspan argues that even if they cannot understand moral reasons as such, typical psychopaths are morally responsible insofar as their behavior can express ill will, but they may be less than fully blameworthy for their moral infractions owing to impairments in behavior control (2016).

[13] See also Wallace 1994; Shoemaker 2007, 2015; Fischer and Ravizza 1998. Paul Litton argues that there is no "meaningful disagreement" between mere rational competence theorists and moral competence theorists, since "the capacity for rational self-governance entails the capacity to comprehend and act on moral considerations" (2008, 351). For arguments that moral responsibility turns on the possession of moral knowledge rather than any particular *capacity*, see Elinor Mason 2017.

[14] Here, Watson follows P.F. Strawson 1962. Watson also raises the case of Robert Harris, a man who callously murdered two innocent teenage boys, but was himself a victim of brutal abuse from a very young age. Watson doesn't identify Harris as a psychopath but uses his case to consider whether some agents "of evil," being unfit for moral dialogue, are inappropriate targets of resentment due to "constraints on moral address" (1987, 268–274).

address" due to their inabilities to recognize moral demands (and the authority of those who address them) as intrinsically reason giving (2011, 309). In his view, psychopaths cannot see our demands that they refrain from harming us as anything but coercive pressures, supplying at best instrumental reasons to comply. Being unable to see the normative force of moral demands, psychopaths are infelicitous targets of resentment.

AMC has been endorsed by a number of theorists, many of whom emphasize the role of emotional capacities in giving uptake to moral address.[15] Some stress the import of being able to feel guilt in response to (negative) reactive attitudes.[16] David Shoemaker focuses on two broader emotional deficits that seem to underlie the psychopath's lack of guilt: his inabilities to care about others and, relatedly, to experience (a certain kind of) empathy. In Shoemaker's view, the capacity to care about others is necessary for being motivated to comply with the reasons exchanged in moral address (2007, 84). He also posits that since the emotional aspect of moral address calls on "the addressee to imaginatively step into the shoes of the other in order to feel what one has put him or her through," moral accountability requires the capacity for "identifying empathy" (2007, 93). Since the psychopath cannot care about the agent who addresses him with resentment, he cannot give the appropriate identifying empathetic response and is thus exempt from moral accountability (ibid; 2015, 146).

The capacities for caring and valuing often play central roles in accounts of morally responsible agency, and not just for proponents of AMC. Antony Duff, for example, argues that moral competence requires "a participant understanding" of at least some values (moral or not) where this involves a "creative capacity to understand the significance of the value ... and to discuss, extend, and criticize its application" (1977, 195). Duff ties understanding the significance of values to emotional sensibilities and a practical commitment to the values in question, where the latter is explained in terms of seeing those values as providing reasons for action. He concludes that owing to deficits in these areas, the psychopath, while intellectually competent, is "seriously defective in practical understanding and rationality," and no more "answerable for his actions ... than ... a young child" (ibid, 199).

Carl Elliott and Grant Gillett take a similar approach, arguing that moral understanding involves the capacities "to create ... one's own moral rules and values," to justify them to oneself and others, and to apply them "imaginatively" by demonstrating insight into the interests of others and one's own weaknesses (1992, 57). Citing

[15] See Scanlon 2008; Smith 2013; Talbert 2008, 2012 for rejections of the argument from moral communication. See Coleen Macnamara 2015 for a detailed response to these challenges.

[16] Macnamara argues that a function of reactive attitudes is securing uptake from their addressees and since "uptake of [resentment] amounts to feeling guilt and expressing it via amends," eligible targets of resentment must be able to feel guilt (2015, 212). Stephen Darwall argues that appropriate addressees of resentment (and its implicit demand) must be assumed to able to "make the same demands of, themselves through acknowledging their validity as in self-reactive attitudes like guilt" (2006, 79). For Darwall, this ability is a matter of competence with the "second-personal reasons" exchanged in moral address, where he describes such reasons as agent-relative reasons "whose validity is grounded in presupposed normative relations between persons" (2006, 78).

certain abnormalities in brain areas associated with higher-order cognitive processing, Elliott and Gillett suggest that the psychopath is unable to adequately integrate his "actions and intentions with his character and commitments to those around him" (ibid, 59–60). Consequently, psychopaths have difficulty forming "stable behaviour patterns as rational and social beings," and this explains their lack of self-regard and their inability to care about morality or other people (ibid, 63).[17]

Notice that Elliott and Gillett describe the psychopath's moral deficits in terms of a broader defect of practical reason. Moral understanding requires the capacity to value, which in turn, requires the capacity for integrated agency, extended over time. This supports a view that has become increasingly popular among responsibility theorists: psychopaths might have diminished moral accountability because they lack the capacities to adequately coordinate their intentions, make realistic, long-term plans, adjust their actions in light of negative outcomes, delay desire satisfaction, and engage in normative self-reflection (Litton 2008; Kennett and Matthews 2009; Levy 2014).[18]

Even those who deny that psychopaths are morally accountable often acknowledge that psychopaths have considerable rational capacities, some of which are morally relevant. Duff, for example, describes psychopaths as adept in areas of practical reason "having to do with a wide range of non-normative beliefs and reasoning" and in "short-term practical reasoning about the satisfaction of desires or impulses" (2010, 209). Watson claims that psychopaths are capable of a complex mode of reflective agency that distinguishes them from mere brutes. They can "get behind" the pain and mischief they cause, and this makes a difference for how we morally respond to them (2011, 316). In light of the considerable abilities that psychopaths do have, some theorists express skepticism about the claims that psychopaths lack the means to acquire moral knowledge or the capacities required for moral competence.[19] In addition, some who deny that psychopaths are "morally accountable," and thus inapt targets of resentment, allow that they might be morally responsible in other senses.[20]

[17] Watson (2013) proposes that the common ground of the psychopath's prudential and moral impairments is an incapacity for a particular kind of normative orientation. Interestingly, Watson's proposal concerns the psychopath's inability to value, where valuing includes "having standards for action, intention, and desire that ... serve as the basis for self-criticism and correction" (2013, 275–276).

[18] Interestingly, Kennett and Matthews (2009) suggest that psychopaths have an impaired capacity for "mental time travel." Neil Levy (2014) takes up this idea, arguing that the psychopath's difficulty with mental time travel obstructs his grasp of moral concepts, such as personhood (and what it means to harm persons), and reduces his moral responsibility.

[19] See, for example, Vargas and Nichols 2007; Brink 2013.

[20] Watson and Shoemaker, for example, both hold that psychopaths may be morally responsible in the "attributability sense," a sense that tracks the relationship between a person's actions and her character (Watson 2011; Shoemaker 2015). Shoemaker delineates a third responsibility category on which psychopaths are sometimes responsible: "answerability," which concerns the agent's ability to "respond to others' demands for justification by citing their judgments about the worth of some reasons over others" (2015, 27). For an argument against the idea that psychopaths (at least as they are often characterized in the philosophical literature) are even attributability-responsible, see Nelkin 2015. For

an argument that, on some descriptions, psychopaths may not even be able to "act for reasons," see Jaworska 2017.

3 Progress and future directions

Having surveyed the relevant literature, we are now well positioned to see what insights we might glean from philosophical treatments of psychopathy and practical reason.

Let's start with a broad observation. In each of the preceding debates, we find some theorists who deem the psychopath "rationally competent" and an opposing group insisting that psychopaths lack the relevant competence, except in a deeply impoverished conception of practical rationality. As I will show, the arguments that develop from this dispute help to illuminate the richness of practical reason – and human agency more broadly – and suggest certain standards for an adequate theory of practical rationality.[1]

The relevant arguments implicate (roughly) four intersecting clusters of abilities that bear on our capacity for practical reason. Some arguments, for example, emphasize our cognitive sophistication. We are complex agents who require intricate coordination, planning, and imagination to successfully identify and pursue the means to our ends. Thus, the capacity for intelligent, goal-directed behavior may not suffice for practical rationality. Attention deficits, disorganized thinking, lack of foresight about the consequences of one's actions, and poor insight into one's own abilities may undermine one's capacity for practical reason.

The second cluster concerns our abilities to experience emotions and to engage in emotional processes. We are not just cognitively sophisticated beings, but we are also emotional beings. What's more, the affective dimension of our psychology is not alienated from practical reason. The debate between sentimentalists and rationalists underscores a now widely endorsed, but sometimes underappreciated, point: the distinction between cognitive capacities and emotional capacities is often nebulous, and even where we can distinguish between them, those capacities often work together to facilitate harmonious deliberation and action. Emotions can help to facilitate access to certain kinds of reasons, clarify reasons, or even serve as practical reasons themselves.[2] Thus, emotional deficiencies might shield certain reasons from view and/or inhibit one's ability to act on such reasons.

[1] Notice here that practical rationality refers to a capacity as opposed to the property that an attitude or act has when compliant with requirements of rationality. For more on this distinction, see John Broome's entry in this volume.

[2] See Patricia Greenspan (2004), along with her entry in this volume, for detailed discussions of the role of emotion in practical reason.

The third cluster concerns our abilities to engage in normative reflection. We can step back and make judgments about our desires, beliefs, and reasons for action. We can guide our behavior in light of those evaluative judgments. We often eschew immediate rewards in favor of pursuing long-term ends that we deem more worthwhile. And we coordinate our intentions and plans accordingly, adjusting course as needed in response to mistakes and new information. The preceding debates invite us to consider how deficits in these areas might disrupt one's agency and interfere with one's ability to make, and to be moved by, normative judgments. If severe enough, such deficits would seem to constitute a considerable defect of practical reason.

The fourth cluster concerns our capacities as valuing agents. We not only make evaluative judgments that guide our actions, but we can engage with value in rich and constructive ways. We can take a "participant stance" that facilitates a more intimate connection with normative material in the world, allowing us to extend and apply our values "creatively." In valuing one's partner or one's career, for example, one comes to see those objects of value as imbued with a special kind of reason-giving force, and this helps us to understand the meaning that others' values have for them. And the same capacities that ground our abilities to care about, and to value, others also seem closely tied to our ability to value ourselves.[3] Finally, further evidence of our dynamic engagement with normative material comes from our ability to bestow value on some object (for example, by loving it) – or, again, to create authority-based reasons for action by exercising normative powers, as we do when we make demands or certain kinds of commitments.[4] Many of the arguments in the preceding debates suggest that the abilities to value in these ways – and to see the interests and authority of others as reason-giving in the relevant sense – are integral to the kind of practical rationality defining of creatures like us.

Taking seriously the multi-layered nature of practical reason has implications for what we should expect from a theory of practical rationality. Minimally, an adequate theory should not stand in tension with our remarkably complex rational natures. And all the better for a theory that helps to explain how specific aspects of our psychology interact to facilitate recognition and responsiveness to reasons for action. The debates canvassed here do not furnish us with a unified theory of practical reason, but they are rife with creative insights that raise interesting questions to keep in mind as we move forward. For example, are certain affective phenomena, such as the dispositions to experience displeasure at cognitive dissonance or regret in response to social censure, preconditions for receptivity to certain kinds of reasons? How, and to what extent, might deficits in fear and sadness obstruct moral learning? What role might positive emotions play in moral reasoning? Given that rational agency and moral competence "go together" in the normal cases, what explains the fact that the psychopath's distinctly moral deficits seem to be far more severe than his (general) rational defects?

[3] For a Kantian approach to this idea, see Christine Korsgaard 1996.

[4] For recent, illuminating work on normative powers, see Ruth Chang 2013.

How do impaired capacities for normative reflection – or again, for valuing – threaten unified agency?

By pondering what may have "gone wrong" with the psychopath, we might gain a better grasp of how practical reasoning ideally operates in rational persons. In this vein, philosophical treatments of psychopathy have laid the ground for considering how other forms of psychopathology might inform moral psychology. For example, given that narcissists seem to value themselves and yet lack the capacity to value others, how do they fare (compared to psychopaths and the general population) as practical reasoners? Do disorders with certain pronounced cognitive impairments, such as attention and memory disorders, tend to have corresponding emotional abnormalities? And if so, how does this combination of features impact moral agency and practical rationality more broadly?

Given that practical reason consists in a suite of abilities that facilitate rational agency, we have a stake not only in identifying those abilities and understanding how they interact but also in determining their respective *significance*. Discussions of psychopathy and practical reason can help to anchor and guide our thinking about this important issue. First, supposing our capacity for practical reason, in part, grounds our statuses as rights-bearers and morally responsible agents, we might wonder which specific abilities are required for the relevant statuses. Relatedly, as the case of the psychopath demonstrates, there are some individuals who have some, but not all, of the relevant abilities. How do we determine whether we should hold such individuals legally responsible for their infractions or, again, whether they have rights that we are bound to respect (such as the right to refuse medical treatment)? And if narrow rational defects can engender more global moral impairment, might our own occasional failures of practical rationality be morally significant in unobvious ways? In attempting to answer these questions, we move beyond (or, perhaps better, expand) the boundaries of practical reason theory, now engaging with theories of responsibility and moral standing, as well as exploring implications for our social, legal, and medical practices.

Mining these debates for insights has raised more questions than it has answered, but there is something to be said for identifying the right questions to ask. The preceding discussion makes plain our interest in acknowledging that practical reason is not a unary capacity but involves a suite of abilities that engage different aspects of our psychology and work together to help constitute us as unified agents. Because they can facilitate rich engagement with others and help secure our passage into the realm of rights and responsibilities, we have a stake in determining how specific capacities might matter for us. As I have argued, the previous debates can help guide our thinking about the nature, function, and significance of practical reason and aid our understanding of the kind of beings we are.

References

Aharoni, E., Sinnott-Armstrong, W. and Kiehl, K. (2012) "Can Psychopathic Offenders Discern Moral Wrongs? A New Look at the Moral/Conventional Distinction," *Journal of Abnormal Psychology*, 121(2): 484–497.

Blair, R. J. (1995) "A Cognitive Developmental Approach to Mortality: Investigating the Psychopath," *Cognition*, 57(1): 1–29.

———. (2008) "The Cognitive Neuroscience of Psychopathy and Implications for Judgments of Responsibility," *Neuroethics*, 1(3): 149–157.

Brink, D. (1989) *Moral Realism and the Foundations of Ethics*. New York: Cambridge University Press.

———. (2013). "Responsibility, Incompetence, and Psychopathy," *Lindley Lecture*, 53: 1–41.

Chang, R. (2013) "Commitment, Reasons, and the Will," in R. Shafer-Landau (ed.) *Oxford Studies in Metaethics*, Vol. 8. Oxford: Oxford University Press, 74–113.

Cleckley, H. (1976) *The Mask of Sanity*. St. Louis, MO: C.V. Mosby Co.

Darwall, S. L. (2006) The Second-Person Standpoint: Morality, Respect, and Accountability*. Cambridge, MA: Harvard University Press.*

Duff, A. (1977) "Psychopathy and Moral Understanding," *American Philosophical Quarterly*, 14(3): 189–200.

———. (2010) "Psychopathy and Answerability," in L. Malatesti and J. McMillan (eds.) *Responsibility and Psychopathy: Interfacing Law, Psychiatry and Philosophy*. Oxford: Oxford University Press, 199–212.

Elliott, C. and Gillett, G. (1992) "Moral Insanity and Practical Reason," *Philosophical Psychology*, 5(1): 53–67.

Fischer, J. M. and Ravizza, M. (1998) *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.

Gill, M. B. (2007) "Moral Rationalism vs. Moral Sentimentalism: Is Morality More Like Math or Beauty?" *Philosophy Compass*, 2(1): 16–30.

Greenspan, P. S. (2003) "Responsible Psychopaths," *Philosophical Psychology*, 16: 417–429.

———. (2004) "Practical Reasoning and Emotion," in A. R. Mele and P. Rawling (eds.) *The Oxford Handbook of Rationality*. Oxford: Oxford University Press, 206–221.

———. (2016) "Responsible Psychopaths Revisited," *Journal of Ethics*, 20: 265–278.

Hare, R. D. (1993) Without Conscience: The Disturbing World of the Psychopaths Among Us. *New York, NY: Guilford Press.*

Jaworska, A. (2017) "Holding Psychopaths Responsible and the Guise of the Good," in S. M. Liao and C. O'Neil (eds.) *Current Controversies in Bioethics*. London, UK: Routledge, 66–77.

Jurjako, M. and Malatesti, L. (2016. "Instrumental Rationality in Psychopathy: Implications from Learning Tasks," *Philosophical Psychology,* 29(5): 717–731.

Kennett, J. (2002) "Autism, Empathy and Moral Agency," *Philosophical Quarterly*, 52(208): 340–357.

———. (2006) "Do Psychopaths Really Threaten Moral Rationalism?" *Philosophical Explorations*, 9(1): 69–82.

———. (2010) "Reasons, Emotion, and Moral Judgment in the Psychopath," in L. Malatesti and J. McMillan (eds.) *Responsibility and Psychopathy: Interfacing Law, Psychiatry and Philosophy*. Oxford: Oxford University Press, 243–260.

Kennett, J. and Fine, C. (2008) "Internalism and the Evidence from Psychopaths and 'Acquired Sociopaths,'" in W. Sinnott-Armstrong (ed.) *Moral Psychology, Volume 3: The Neuroscience of Morality: Emotion, Brain Disorders, and Development*. Cambridge, MA: MIT Press, 173–190.

Kennett, J. and Matthews, S. (2008) "Normative Agency," in K. Atkins and C. Mackenzie (eds.) *Practical Identity and Narrative Agency*. New York, NY: Routledge.

———. (2009) "Mental Time Travel, Agency and Responsibility," in M. Broome and L. Bortolotti (eds.) *Psychiatry as Cognitive Neuroscience: Philosophical Perspectives*. Oxford: Oxford University Press, 327–349.

Kiehl, K. and Buckholtz, J. W. (2010) "Inside the Mind of a Psychopath," *Scientific American Mind*, 21(4): 22–29.

Korsgaard, C. (1996) *The Sources of Normativity*. Cambridge: Cambridge University Press.

Kumar, V. (2016) "Psychopathy and Internalism," *Canadian Journal of Philosophy*, 46(3): 318–345.

Levy, N. (2014) "Psychopaths and Blame: The Argument from Content," *Philosophical Psychology*, 27(3): 351–367.

Litton, P. (2008) "Responsibility Status of the Psychopath: On Moral Reasoning and Rational Self-Governance," *Rutgers Law Journal*, 39(349): 350–392.

Macnamara, C. (2015) "Blame, Communication, and Morally Responsible Agency," in R. Clark, M. McKenna, and A. Smith (eds.) *The Nature of Moral Responsibility*. Oxford: Oxford University Press, 211–236.

Maibom, H. L. (2005) "Moral Unreason: The Case of Psychopathy," *Mind and Language*, 20(2): 237–257.

———. (2010) "Rationalism, Emotivism, and the Psychopath," in L. Malatesti and J. McMillan (eds.) *Responsibility and Psychopathy: Interfacing Law, Psychiatry and Philosophy*. Oxford: Oxford University Press, 227–242.

Mason, E. (2017) "Moral Incapacity and Moral Ignorance," in R. Peels (ed.) *Perspectives on Ignorance from Moral and Social Philosophy*. Oxford: Oxford University Press.

McGeer, V. (2008) "Varieties of Moral Agency: Lessons from Autism (and Psychopathy)," in W. Sinnott-Armstrong (ed.) *Moral Psychology, Volume 3: The Neuroscience of Morality: Emotion, Brain Disorders, and Development*. Cambridge, MA: MIT Press, 227–258.

Nelkin, D. K. (2015) "Psychopaths, Incorrigible Racists, and the Faces of Responsibility," *Ethics*, 125(2): 357–390.

Nichols, S. (2002) "How Psychopaths Threaten Moral Rationalism," *The Monist*, 85(2): 285–303.

———. *(2004)* Sentimental Rules: On the Natural Foundations of Moral Judgment. *Oxford: Oxford University Press.*

Prinz, J. (2006) "The Emotional Basis of Moral Judgments," *Philosophical Explorations*, 9(1): 29–43.

———. (2007) *The Emotional Construction of Morals*. Oxford: Oxford University Press.

Roskies, A. (2003) "Are Ethical Judgments Intrinsically Motivational? Lessons From 'Acquired Sociopathy,'" *Philosophical Psychology*, 16(1): 51–66.

———. (2008) "Internalism and the Evidence from Pathology," in W. Sinnott-Armstrong (ed.) *Moral Psychology, Volume 3: The Neuroscience of Morality: Emotion, Brain Disorders, and Development*. Cambridge, MA: MIT Press, 191–206.

Scanlon, T. M. (1998) *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.

———. (2008) *Moral Dimensions: Permissibility, Meaning, Blame*. Cambridge, MA: Harvard University Press.

Shoemaker, D. (2007) "Moral Address, Moral Responsibility, and the Boundaries of the Moral Community," *Ethics*, 118(1): 70–108.

———. (2015) *Responsibility from the Margins*. Oxford: Oxford University Press.

———. (2014) "Do Psychopaths Refute Internalism?" in S. Thomas (ed.) *Being Amoral: Psychopathy and Moral Incapacity*. Cambridge, MA: MIT Press, 187–208.

Smith, A. (2013) "Moral Blame and Moral Protest," in D. Justin Coates and N. Tognazzini (eds.) *Blame: Its Nature and Norms*. Oxford: Oxford University Press, 27–48.

Smith, M. (1994) *The Moral Problem*. Malden, MA: Blackwell.

———. (2008) "The Truth about Internalism," in W. Sinnott-Armstrong (ed.) *Moral Psychology, Volume 3: The Neuroscience of Morality: Emotion, Brain Disorders, and Development*. Cambridge, MA: MIT Press, 207–216.

Strawson, P. F. (1962) "Freedom and Resentment," in G. Watson (ed.) *Proceedings of the British Academy, Volume 48: 1962*. Oxford: Oxford University Press, 1–25.

Talbert, M. (2008) "Blame and Responsiveness to Moral Reasons: Are Psychopaths Blameworthy?" *Pacific Philosophical Quarterly*, 89(4): 516–535.

———. (2012) "Moral Competence, Moral Blame, and Protest," *The Journal of Ethics*, 16(1): 89–109.

Turiel, E. (1983) The Development of Social Knowledge: Morality and Convention. *Cambridge: Cambridge University Press.*

Vargas, M. and Nichols, S. (2007) "Psychopaths and Moral Knowledge," *Philosophy, Psychiatry, and Psychology*, 14(2): 157–162.

Wallace, R. J. (1994) *Responsibility and the Moral Sentiments*. Cambridge, MA: Harvard University Press.

Watson, G. (1987) "Responsibility and the Limits of Evil: Variations on a Strawsonian Theme," in F. Schoeman (ed.) *Responsibility, Character, and the Emotions*. Cambridge: Cambridge University Press, 256–286.

———. (2011) "The Trouble with Psychopaths," in *Reasons and Recognition: Essays on the Philosophy of T.M. Scanlon*. Oxford: Oxford University Press, 307–323.

———. (2013) "Psychopathic Agency and Prudential Deficits," *Proceedings of the Aristotelian Society*, 113: 269–292.

19. Practical reason and social science research[1]

[1] The chapter draws on work published previously in Tiberius 2015, 2016. The authors would like to thank Routledge Publishing Inc. and Kelly James Clark for their permissions.

Valerie Tiberius and Natalia Washington

In many areas of philosophy, it is becoming more and more mainstream to appeal or at least refer to social science research. For example, in moral psychology, the empirically informed approach is well established in the literature on moral judgment, moral emotions, and moral responsibility (Greene, 2013; Nichols, 2004; Prinz, 2007; Kelly, 2011; Doris, 2016; Roskies, 2006; Vargas, 2013). Does work in the social sciences have any bearing on philosophical questions about practical reason or reasoning? While there has been some excellent work that explores this question (e.g., Stein, 1996; Bishop and Trout, 2004), the growing empirical wave is less noticeable here. But this may not be because empirical research is irrelevant to the philosophical questions in the area. In this chapter, we will consider three questions about practical reason or reasoning that might be illuminated by attention to research in the social sciences. First, we'll consider whether research on the role of sentiments on moral judgment is relevant to the philosophical debate between moral rationalism and moral sentimentalism. Second, we'll consider whether research on the unconscious causes of action undermines philosophical views about the relationship between reasons and actions. Third, we'll consider whether research on implicit bias is relevant to the normative question of how we can reason better. Along the way, we'll see that when philosophical arguments and theories about practical reason make empirical assumptions – which they often do – empirical research can help to test these assumptions. We'll also see that philosophical theories don't always make the assumptions they are accused of making by their critics. Finally, we'll see that we can make better prescriptions for how to reason better if we have a better sense of what we're doing wrong.

1 Rationalism and sentimentalism about moral judgments

Let's define metaphysical rationalism about moral judgment as the view that the truth of a moral judgment is determined by rational principles.[1] Rationalists also hold what we might call epistemic rationalism: the claim that reasoning is crucially involved in arriving at moral judgments. The two views are connected, because if epistemic rationalism is false, then there is no connection between the activity of justifying a moral judgment and the truth, which makes metaphysical rationalism unappealing. What does the empirical evidence tell us about this picture?

Emotions influence and cause moral judgments

We can start with evidence for sentimentalism, which is often taken to be evidence against (epistemic) rationalism. According to sentimentalism, emotions play an essential role in the making of moral judgments, and reasoning plays a much less central role or no role at all. Sentimentalists differ over whether moral judgments are reports about or expressions of our sentiments and over which sentiments are crucial to moral judgment and under what conditions. Despite these disagreements, sentimentalists all agree that moral judgments are not made true by rational principles (so, they reject metaphysical rationalism). If moral judgments are reports about our emotions, then they are made true by facts about our emotions (perhaps facts about the emotional response we would have in certain conditions). For example, in Prinz's theory, the moral judgment "lying is wrong" is true (if it is true) in virtue of the sentimental disposition of the person making the judgment; it is not made true by the Categorical Imperative. Alternatively, if moral judgments are expressions of our emotions, then whatever story might be told about whether and in what sense moral judgments are truth-apt, it couldn't be a story that appeals to rational principles (Blackburn, 1984; Gibbard, 1992). Sentimentalism of either form presents a picture of moral judgment that is at odds with the rationalist picture, according to which moral judgments do sometimes (when true) track the deliverances of reason.[2] We could put the point this

[1] The Kantian rationalist also holds that moral judgments give us reasons that motivate us insofar as we are rational, independently of our non-rational sentiments or desires. Empirical evidence is likely to bear on this claim, but we do not consider it here.

[2] Things get more complicated for certain sentimentalists, D'Arms and Jacobson (2000), for ex-

way: if epistemic rationalism is true and sentiments, not reasoning, are driving our moral judgments, then metaphysical rationalism becomes unappealing, because our actual judgments are not tracking rational principles.

The case for sentimentalism, particularly the case defended by Jesse Prinz (2007), often starts with evidence that emotions both influence and cause moral judgment (evidence for what we have called epistemic rationalism). There is a good deal of evidence for the influence of emotions on moral judgment. To give one colorful example, here's an experiment that shows the effect of disgust on moral judgment. Psychologists asked participants to answer questions about the moral propriety of four different scenarios, two having to do with incest between first cousins, one having to do with the decision to drive rather than walk to work, and the last having to do with a studio's decision to release a morally controversial film (Schnall et al., 2008). The participants were divided into three different groups: no-stink, mild-stink, and strong-stink, distinguished by the amount of stink – in the form of "commercially available fart spray" sprayed into a nearby trash can – in the environment. The results of the experiment were that feelings of disgust increase people's tendency to make harsh moral judgments. Other experiments have shown that anger makes people more punitive and harsh in their moral judgments about crimes against persons (Lerner, Goldberg, and Tetlock, 1998; Seidel and Prinz, 2013).

There is also evidence that emotions cause us to make moral judgments that we would not otherwise make. For example, Thalia Wheatley and Jonathan Haidt hypnotized half the participants in one study to feel disgust when they heard the word "often" and the other half to feel disgust when they heard the word "take." All the participants then read some scenarios, one of which was this one:

> Dan is a student council representative at his school. This semester he is in charge of scheduling discussions about academic issues. He {tries to take/ often picks} topics that appeal to both professors and students in order to stimulate discussion.
>
> (Wheatley and Haidt, 2005)

Participants were randomly assigned to a group that got one or the other of the phrases in the square brackets; no student saw both the phrases in the brackets.

To those of us who have not been hypnotized, it doesn't seem like Dan has done anything wrong. Moreover, participants who read the scenario that did not contain their disgust-inducing word did not rate Dan's behavior as wrong. However, for the students who did feel disgust (because they read the scenario with the word that induced disgust in them), there was a tendency to rank Dan's actions as wrong. This

---

ample, who hold that value judgments are made true by facts about reasons for emotions, where reasons are considerations that bear on fittingness. These reasons for emotions are not principles of reason that are independent of our sentimental nature, however, so we maintain that all sentimentalists reject

is a case in which the people in question would not have made a judgment of moral wrongness at all were it not for the emotion of disgust they experienced.

The fact that emotions influence moral judgments does not establish that moral judgments *are* emotional responses, nor even that emotions are an essential part of moral judgment. (Nor does Prinz think it does – he offers this evidence as part of a larger case.) This would only be an argument for rejecting (epistemic) rationalism if the sentimentalist understanding of moral judgment were the only way to explain the influence of emotions on moral judgment. Other explanations are possible; it could be that emotions influence moral judgment in the way that wearing rose-colored glasses can influence your judgment about the color of the sky: the glasses sway your judgment, but they're not part of the content of the judgment that the sky is pink. Evidence that emotions cause moral judgments is somewhat more difficult to explain away, but the person who rejects sentimentalism can still argue that when moral judgments are entirely caused by emotions, they are akin to manipulated illusions; after all, it has not been shown that all of our moral judgments are such that we would not make them were it not for our emotions.

Perhaps the sentimentalist would have a stronger argument if there were evidence that we simply cannot make moral judgments without emotions. We'll consider this possibility in the next section.

Emotions are necessary for moral judgment

Some have thought that psychopaths provide evidence that we cannot make moral judgments without emotions, because (to oversimplify) psychopaths are amoral and they do not experience normal emotions like sympathy or compassion. If we cannot make moral judgments without emotion, it might seem like emotions are essential to moral judgment in a way that supports the sentimentalist characterization.

Psychopathy is a personality disorder characterized by lack of empathy, impulsivity, egocentrism, and other traits. The disorder is often diagnosed by the Psychopathy Checklist, which asks a number of questions that cluster under the headings "aggressive narcissism" and "socially deviant lifestyle" (Hare, 2003). Because psychopaths lack empathy, they are of interest to those who think emotions like empathy are essential to moral judgment. Shaun Nichols (2010), for example, thinks that the evidence from psychopathy counts against rationalism because psychopaths do not have defects of reasoning and yet do not seem to make moral judgments in the same way that the rest of us do. The basic argument goes this way:

1. Psychopaths do not make a distinction between moral wrongs and conventional wrongs.

---

metaphysical rationalism.

2. It is the defect to the emotional response system that is responsible for psychopaths' decreased ability to distinguish moral wrongs from conventional wrongs.

3. Therefore, a functioning emotional response system is essential to moral judgment.

The conclusion of this argument is taken to be strong evidence for sentimentalism (and against epistemic rationalism).

Let's consider the steps of this argument in detail. The first thing to notice is the importance of the distinction between "moral" and "conventional." Conventional norms, such as "you shouldn't go outside in your pajamas," are different from moral norms in a variety of ways. Moral norms are thought to be more serious and to have wider applicability than conventional norms. Conventional norms are thought to be contingent on an authority (such as a teacher or the law or, in the case of the pajamas, a culture), and they receive a different kind of justification from moral norms, which are often justified in terms of harm or fairness (Nichols, 2004). For example, young children will say that it would be wrong to pull another child's hair, even if the teacher said it was OK, because pulling hair hurts, whereas, the wrongness of chewing gum in class depends on the teacher's forbidding it.

It is widely believed that psychopaths don't really understand this distinction (Blair, 1995); that is, psychopaths think of what's morally wrong as what's prohibited by the local authority, and they do not see moral transgressions as more serious than other kinds of violations of rules. This claim is now considerably more controversial than it used to be, but even the latest research confirms that the "affective defect" part of psychopathy does predict poor performance in distinguishing moral from conventional wrongs (Aharoni, Sinnott-Armstrong, and Kiehl, 2012). Because they don't feel bad when others suffer, "they cannot acquire empathetic distress, remorse, or guilt. These emotional deficits seem to be the root cause in their patterns of antisocial behavior" (Prinz, 2006). Further, these emotional deficits seem to be responsible for the fact that they don't make the same kinds of moral judgments that the rest of us do.

Others, however, argue that psychopaths have defects of rationality that play a crucial role in their moral defects, and hence that they do not provide an obvious problem for rationalism. According to Jeanette Kennett (2006), the evidence suggests that psychopaths have at best a weak capacity to stand back from and evaluate their desires, to estimate the consequences of their actions, to eschew immediate rewards in favor of longer term goals, to time order, to resolve conflicts among their desires, to find constitutive solutions. To these rational shortcomings we may add that psychopaths frequently choose grossly disproportionate means to their immediate ends or fail to adopt the necessary means to their proclaimed ends (Kennett, 2006, p. 77).

The evidence for sentimentalism from psychopathy, then, is suggestive but not conclusive. First, psychopathy is a tricky category that includes multiple defects, not all

of which are emotional, and it is a complex matter to figure out what is responsible for what. Second, even if we agree that what is important is the ability to distinguish moral from conventional wrongs, and even if we agree that psychopaths with emotional defects are thereby hindered in their ability to make this distinction, we have not shown that reasoning has no important role in the making of moral judgments.

Reasoning does not play the role that the rationalist thinks it does

Notice that even if the previous evidence for the view that emotions are essential to the making of moral judgments were conclusive, this would not rule out the possibility that reasoning plays some role in moral judgment, nor that moral judgments are justified by rational principles. After all, if the rationalist's primary claim is about what makes moral judgments true (metaphysical rationalism), she can be agnostic about what mechanism is operating when we make moral judgments. In other words, the rationalist could accept that a properly tuned emotional system is necessary for making appropriate moral judgments while maintaining that what is distinctive about moral judgment is that they are justified by rational principles (Kennett, 2006). That said, the metaphysical rationalist will not be in a good position if it turns out that reasoning plays *no* role in moral judgment: if this were the case, then there would seem to be no connection between what makes our moral judgments true and the process that we use to justify them. To examine this possibility, we turn to a further attack on rationalism.

Psychologist Jonathan Haidt (2001) argues that the role of reasoning in moral judgment is to provide post hoc rationalizations of the judgments we come to on the basis of sentiment and affect. While it is possible for us to reason about our moral judgments, according to Haidt, this happens rarely. He calls his theory of moral judgment "the social intuitionist model" (SIM), because moral judgments are quick and intuitive, and when reasoning is used, it is usually social reasoning of the kind that takes place as people talk and argue with each other to try to figure things out. SIM does allow that individual reasoning or "private reflection" occurs and can affect our judgments but maintains that it is not the typical cause of moral judgment. If true, this could put some pressure on the rationalist theory, because it would mean that our moral judgments are not responsive to reasoning nor to the rational principles that reasoning would track. This would make rational principles seem otiose so that any argument for their role in the justification of moral judgments would have no applicability to actual human beings.

We can't review all of Haidt's evidence here, so we will focus on one piece that has attracted a great deal of attention from philosophers. This is the phenomenon of moral dumbfounding. Haidt first points out that we can (and often do) confabulate

justifications for intuitive judgments that were not made by reasoning. This creates the illusion of objective reasoning when what is really happening is post hoc rationalization. Moral dumbfounding happens when a person cannot find any reasons for the moral judgment she makes and yet continues to make it anyway. In the study that introduced the phenomenon, subjects are were presented with the following scenario:

> Julie and Mark, who are brother and sister, are traveling together in France. They are both on summer vacation from college. One night they are staying alone in a cabin near the beach. They decide that it would be interesting and fun if they tried making love. At very least it would be a new experience for each of them. Julie was already taking birth control pills, but Mark uses a condom too, just to be safe. They both enjoy it, but they decide not to do it again. They keep that night as a special secret between them, which makes them feel even closer to each other. So what do you think about this? Was it wrong for them to have sex?
>
> (Bjorklund, Haidt, and Murphy, 2000)

Most subjects say that the siblings' behavior is wrong, and they offer reasons for their judgment. They say Mark and Julie may have a deformed child, or that it will ruin their relationship, or cause problems in their family, and so on. But because of the way the scenario is constructed, the interviewer can quickly dispel their reasons, which leads to the state of dumbfounding. According to Haidt, in an interview about his findings, dumbfounding only bothers certain people:

> For some people it's problematic. They're clearly puzzled, they're clearly reaching, and they seem a little bit flustered. But other people are in a state that Scott Murphy, the honors student who conducted the experiment, calls "comfortably dumbfounded." They say with full poise: "I don't know; I can't explain it; it's just wrong. Period."
>
> (Sommers, 2005)

Dumbfounding, the argument goes, shows that most people don't make moral judgments for reasons and that moral judgments are usually not responsive to reasoning. Rather, people offer post hoc rationalizations of their emotional convictions, and when these rationalizations are undermined, they stick with their convictions anyway. Does the evidence show that our moral judgments are not responsive to reasoning and hence that rational principles are otiose?

First, let's consider whether Haidt and the rationalists mean the same thing by "reasoning." If they do not, then Haidt's challenge won't necessarily undermine rationalism. Haidt's picture of moral reasoning seems rather different from what Kantians – some of the main proponents of rationalism in moral philosophy – take moral reasoning to be. Haidt concedes rare cases in which people "reason their way to a judgment

by sheer force of logic" (Haidt and Bjorklund, 2008, p. 819). But this is a caricature: moral rationalists do not typically think that we reason ourselves into moral positions by sheer force of logic. One tool of moral reasoning that Kantians think is particularly important is universalization: when we're unsure what to do, we should ask ourselves whether the intention of our action requires making a special exception for ourselves or whether it is an intention that we think is acceptable for everyone to have. This sort of reasoning is not the sheer force of logic.

Second, the rationalist need not hold that moral judgments are usually or even typically responsive to reasoning. It is open to them to hold that the judgments we make are often unjustified as long as we are capable of reasoning and capable of changing our judgments in response to that reasoning. The research does not show that we lack these capacities, and indeed Haidt seems to think that we do have them (given a more generous interpretation of what reasoning amounts to). After all, Haidt concedes that we sometimes arrive at judgments through private reasoning. He also thinks that we engage in social reasoning – reasoning with each other in the form of argument and gossip – that does change our moral judgments. Rationalists do not need to assume that moral reasoning is always private. Indeed, reasoning with each other might help us overcome our biases so that we can be more impartial and better universalizers.

Does the empirical evidence really provide a fundamental challenge to rationalism? What does seem to be threatened is a picture according to which we always arrive at our moral judgments by engaging in rational reflection and are then motivated to act on these judgments by the sheer recognition of their rational status. It's unlikely that even Kant held this extreme view. Ultimately, whether Kantian reasoning can *justify* our moral judgments depends on some deep issues in metaethics. In particular, it depends on whether there really are any rational principles that provide a foundation for our moral reasons. This is one of the fundamental philosophical disputes between the sentimentalist and the rationalist. But this debate ultimately concerns philosophical questions about the nature of practical reasons, not psychological facts about the causes of moral judgment. The empirical challenge to epistemic rationalism is on better footing, but how challenging this challenge really is depends on what the rationalist thinks about the possible role for emotion in moral judgment. As we've seen, even the epistemic rationalist need not deny that emotions have some role. As far as the empirical challenge goes, then, the door for rationalism is still open.

2 Reasons, action, and agency

Historically, many philosophers have identified our rational capacities as the most important human abilities. Our capacities to reason about what to do and to act for the reasons we arrive at in deliberation have at various points been taken to be the criterion for moral agency, the key to our flourishing, and the basis for responsible agency. John Doris (2016) argues that "reflectivism" – the idea that "the exercise of human agency consists in judgment and behavior ordered by self-conscious reflection about what to think and do ... [and the view that] the exercise of human agency requires accurate reflection" (19) – is also an assumption of the standard philosophical conception of a person.

Recent social science research raises some worries about these assumptions. Indeed, Doris argues for skepticism about reflective agency on the basis of psychological research that shows that many of our actions and cognitions are caused by non-reflective processes that would not be endorsed as reasons for those actions and cognitions upon reflection. What is this evidence and what does it show? Should philosophers whose favored theories assume that we're fairly rational be worried?

Recent challenges to virtue ethics have made familiar a program in social science research that reveals how our actions are often influenced by situational factors that we would not endorse as reasons for acting (Doris, 2002). For example, the bystander intervention effect, which has been observed in a number of different contexts, is that people are less likely to help someone in need when there are other people around who are not doing anything (Latane and Darley, 1970). Surely people don't explicitly think that the fact that there are others around is a reason not to help someone; rather, they are influenced subconsciously by irrelevant factors. In the infamous Milgram experiments, people essentially tortured their fellow human beings, responding to the authority of the psychologists in charge, rather than to the reasons against inflicting excessive pain on an innocent person that most of them recognized. It is unlikely that the people who initiated the shocks would endorse "obedience to authority" as a reason to risk killing an innocent person, and no one predicted that people would be influenced by that factor so strongly. In the Good Samaritan study (Darley and Batson, 1973), people acted on considerations of punctuality rather than on the reasons for helping others that many of them were about to preach. Again, people in this experiment were not making a conscious decision to be punctual; rather, they were influenced unconsciously by their feelings of being in a hurry, and they were influenced in ways they would likely not endorse if they thought about it. In all of these cases, and many other examples from social psychology, people would likely try to counteract the influence

of these unrecognized situational influences if they knew about them, suggesting that they are not acting on reasons they endorse.

There is other evidence of our irrationality that doesn't have to do with virtue and moral behavior. A number of studies suggest that our choices of consumer products are not made for the reasons we think they are. These studies also tend to show that we confabulate reasons after we choose so that our choices seem to have been made for considerations we endorse as reasons, even though they really weren't. In one such study, people were asked to choose among four pairs of stockings that were (unbeknownst to the shoppers) exactly the same. People tended to choose the stockings on the right, but they explained their choice by referring to the better quality of the stockings they chose (Wilson and Nisbett, 1978). Other, more elaborate, studies show that analyzing our reasons tends to change our attitudes toward political candidates, our beliefs about whether our romantic relationships will last, and our judgments about how much we like different posters (Wilson, Kraft, and Dunn, 1989; Wilson and Kraft, 1993; Wilson et al., 1993). In this last study, participants were asked to evaluate two types of posters: reproductions of impressionist paintings and humorous posters, such as a photograph of a kitten perched on a rope with the caption, "Gimme a Break" (Wilson et al., 1993). All of the participants were asked to rate how much they liked each poster and then allowed to choose a poster to take home. The reflectors were instructed to write down their reasons for feeling as they did about the posters before giving their ratings, while the controls did a cognitive task not related to reflecting on reasons (a filler task). The results of the study were that the reflectors rated the humorous posters significantly higher than the controls did and were much more likely to take these humorous posters home. A few weeks later, the researchers telephoned all the participants and asked them several questions about the posters they had chosen (how much they liked them, whether they still had them, whether they had hung them up on their walls). The reflectors were less satisfied with their posters than the people who did not reflect on their reasons before choosing a poster.

Looking at this research, it seems like we don't often have concrete reasons that we use as a basis for deriving our attitudes, beliefs, and judgments (such as their poster preferences). When we are asked to analyze our reasons, we just make up something that's easy to think of or that we believe will make sense to other people (that is, we confabulate). These confabulations then lead us away from the attitudes, beliefs, and judgments we made before we started thinking about it (such as a preference for the impressionist poster). There's nothing necessarily wrong with this – analyzing your reasons might improve your beliefs after all. The point for our purposes is that people think their beliefs and judgments really are based on the reasons they offer when asked about their choice, but this is mistaken. If this is how we are, then the picture of us as competent rational agents, deliberating about our reasons and then acting on the results of our deliberation, seems to be a bit tarnished. Further, the picture of us as responsible agents is threatened as long as we assume (as many do) that we are responsible only for what we rationally choose to do.

But how tarnished, really? Is it true that we do not have the rational capacity to reflect on, endorse, and act for reasons that is required by responsibility, according to many compatibilists (e.g, Fischer and Ravizza, 2000)? The evidence we have surveyed is evidence that we don't use this capacity all the time, but it isn't evidence that we don't have it. If we sometimes act for the reasons we think we do – if we sometimes reflect on, endorse and act for the reasons we recognize as reasons, even though not always – this will be enough to show that we are at least sometimes responsible agents.

Importantly, the psychologists who have done the research just presented as evidence against the effectiveness of our rational capacities do not tend to endorse any strong claim about our lacking these capacities altogether. As their research has continued, they have found that there are some variables that seem to make us better at knowing our reasons. For instance, the more knowledgeable we are about something, the more we understand our own reasons for making choices with respect to that thing (Halberstadt and Wilson, 2008). Someone who was an expert in stockings would probably have seen that the stockings were identical (as a few people in the actual study did) and would not have fabricated reasons for choosing one over another. Someone who was able to articulate what it is about impressionist art that makes it beautiful might have chosen the art poster rather than the kitten poster after reflecting on her reasons for preferring one to the other. Knowledge is not all powerful, of course, but it does sometimes help, which is evidence that we should not be so pessimistic as to think we never do things for the reasons we think we do.

Further evidence that reflective processes (such as conscious reflection on values) matter comes from a fairly well-established literature on the efficacy of reflective goal-setting exercises on the attainment of goals (Locke and Latham, 2002; Covington, 2000; Ellis et al., 2014). These studies show that engaging in a reflective exercise in which subjects identify goals, articulate strategies for achieving them, and describe their ideal future has significant effects on goal achievement (measured by GPA, course load, and graduation rates). The goal-setting exercise is done alone online by participants. In one set of studies involving college students, the exercise has eight parts (Morisano et al., 2010). Step 1 is designed to get students thinking about their ideal futures, step 2 asks them to identify "clear and specific goals" they could set to reach this ideal future, step 3 asks students to evaluate the goals, step 4 asks them to think about the impact that achieving the goals would have on them and on others, steps 5–7 encourage more detailed planning about how to meet these goals (including identifying sub-goals), and in step 8 students are encouraged to think about how much they are committed to the goals. We take these goal-setting exercises to be reflective in the relevant sense (recall Doris's definition of reflectivism as the view that "the exercise of human agency consists in judgment and behavior ordered by self-conscious reflection about what to think and do"). Compared to the control group, students who performed the reflective goal setting exercise achieved increased GPA, were more likely to complete a full course load, and had a reduced tendency to report negative affect. Since almost all of the students in the study had academic goals, it seems that by thinking about their goals

and how to achieve them, they were able to better achieve them. A similar study with a similar intervention showed improvements in academic performance and drop-out rates (Schippers, Scheepers, and Peterson, 2015).

Finally, we may act rationally without explicitly acting for the reasons we endorse in the instant. To see this, consider Andy Clark's (2007) helpful notion of "ecological control". Ecological control is the kind of goal directed effort served by the creation, cooptation, maintenance, and use of subpersonal and non-biological resources in the surrounding environment. Rather than micro-managing every individual task, ecological control allows us act fluently and efficiently in the moment – think of practicing a difficult golf swing; creating automatic reminders of loved ones' birthdays; or removing oneself to a quiet, snack-free environment to finish a paper. Ecological control comprises two senses of rational action. You can take ecological control (sense 1) by, for instance, entering important birthdays and programming reminders in your digital calendar. This puts a tool into place that you can rely on at a later date in an exercise of ecological control (sense 2) (Holroyd & Kelly, 2016). In the first sense, you are co-opting your environment, using your abstract reasoning skills to structure your ecological niche. In the second sense, you are relying on previously deployed structures in order to bring your behavior in line with your goals without the need for explicit reflective effort. Ecological control (in both senses) is important for two reasons. First, it is rational insofar as it increases our capacity to act consistently with the reasons we endorse overall. Second, ecologically controlled actions are not undermined by the psychological evidence, insofar as we are not required to bring the reasons we endorse to mind for each individual action. Indeed, once we recognize the ways in which situations influence us, ecological control allows us to choose situations that will make us better able to act in accordance with on the reasons that we recognize as good reasons for doing things.

In our view, then, the reasonable thing to conclude from the evidence is not terribly exciting: we are sometimes rational, but not as often as we thought we were. How bad a result this is for philosophical theories depends on exactly what assumptions the theory makes about the psychological processes that distinguish agency from non-agency. In general, however, we can say that psychological research does not undermine deeply held philosophical convictions such as the idea that there are real psychological differences between people who are forced or hypnotized to do things and people who choose to do them and between people who are incapable of understanding what a moral reason is and people who just ignore moral reasons out of selfishness or malice. Exactly how to characterize these differences compatibly with our actual psychological capacities is a matter for future debate, but it seems safe to assume that we are at least to some degree creatures who can grasp reasons and choose to act on them.

In our final section, we want to examine a case in which our actions are caused by forces we do not endorse as reasons that is of particular interest to the philosophical community. Our suggestions about what to do about this will illustrate the ways in which we can claim agency despite not being perfectly rational creatures.

3 Good reasoning and overcoming bias

As we have seen, research in the human sciences can reveal ways in which we are likely to reason poorly or likely to be influenced by factors which we do not endorse as reasons. The influence of disgust on moral judgment, the moral dumbfounding effect, and situational influences on helping behavior are all good examples of this. One might wonder, given this influx of information, what our response as individuals who are motivated to reason well should be. Now that we know, what do we do?

In order to address this question, let's look closely at one problem that is particularly important in the context of academic philosophy. It is well-known today that women are underrepresented in academic philosophy, both in terms of publication and tenure track positions. While there are likely to be many heterogeneous causes for this phenomenon (Antony, 2012), one explanation that is gaining traction is that this underrepresentation is partially driven by an implicit association many people have between maleness and academic achievement (Saul, 2017). Briefly, this association may contribute to biased judgments about who is and isn't a good philosopher – especially since academic philosophy, in comparison to other disciplines, tends to emphasize intangible qualities like 'brilliance' or 'raw talent' (Leslie et al., 2015). So, for example, even though a graduate admissions committee member may explicitly believe that women and men are equally qualified to study philosophy, subconscious cues about gender in writing style may lead them to choose more writing samples from candidates they perceive as male among equally qualified applications.

The problem of implicit bias – unconscious negative evaluations of others based on race, sex, gender, and so on – is a good case study for the problem of implicit cognition generally. Taken alongside evidence from the heuristics and biases tradition (Kahneman, 2011) and dual systems theory (Haidt, 2001; Greene, 2009), the picture of ourselves that is emerging is one according to which we are less often the authors of our own judgments and actions than we would hope. According to this new view, much of our everyday behavior is driven by hidden features of our psychology that operate automatically and without conscious control, the causes or triggers of which are unavailable to introspection. Worse, these mechanisms can influence us in ways that we would not endorse upon reflection. From the perspective of the practical reasoner, this problem may seem daunting. We'll confront it in two steps: First, we'll look at the tricky characteristics of these mechanisms as they are illuminated in the academic philosophy example, and at strategies for dispelling or mitigating implicit bias which take these

characteristics into account. Second, we'll draw forth a few broader lessons about practical reasoning and consider the perhaps counterintuitive implication that many of the most effective prescriptions for better practical reasoning emphasize collaborative, institutional, or otherwise anti-individualist methods.

Hidden mechanisms

What happens in a case where an explicitly egalitarian reader's judgments about which pieces of philosophical writing are (say) at the graduate level are unconsciously influenced? The first thing to note about this case is that evaluating philosophy papers is a cognitively demanding, temporally extended reasoning process. Our reader is likely to consciously consider reasons she has for putting particular writing samples in the 'accept' pile – for example, because the paper is organized around a valid argument about a philosophically interesting topic. She may even have a written rubric or checklist. Unfortunately, cues in writing style of which our reader is unaware bias her judgments in a way that she would not endorse if it were brought to her attention. Here are three features of this reasoning process worth highlighting:

1. Implicit cognition is often a partial cause – Even if the reader knows about implicit cognition in general, she does not feel the influence of her bias as it is triggered. She therefore does not know whether or how implicit mechanisms are a factor in a particular decision. Thus, although her judgments about which papers to accept are for the most part driven by her reflective capacities, they are infected in a way she is not able to recognize. Implicit social cognition does not bypass reflection but is, rather, a pernicious influence on it.

2. Implicit cognition is obscured by the introspection illusion – Our reader has reliable access to her explicit, reflective reasons for choosing particular papers ... she can simply look down at her checklist and notes! Further, due to her commitment to egalitarianism, she is likely to deny that the gender of the author was a consideration in her decision-making. Our reader's self-reports, therefore, will match neither the causes of her behavior, nor its results.

3. Implicit cognition is deeply rooted – Being made aware that bias is or is likely to be expressed in the course of her reasoning does not allow the reader to control or suppress its influence through an act of will. Such efforts are likely to backfire. Ridding herself of her biases is similarly difficult. Indeed, the notion that implicit bias is a neatly encapsulated component of one's psychological makeup that can be 'removed' in the same way as a kidney from a body or a fuel pump from a car is not likely to be validated. An objectionable association between maleness and academic achievement may be deployed by an implicit mechanism or mechanisms with a much more general domain.

What, then, can our reader do next time in order to make fair judgments according to her egalitarian commitments? There are several interesting research programs focused on interventions for implicit bias, though most are in their nascent stages (for up-to-date information, see Brownstein, 2015). We'll consider just a few examples. First, because implicit cognition is a partial cause of her ultimate judgments, our reader must recognize that spending more time considering her explicit reasons is unlikely to be of any help. Rather, if possible, she should find a way to prevent her biases from being triggered in the first place. This is why much has recently been made of the efficacy of blind evaluations. Names carry demographic information, so removing them is a type of environmental scaffolding that eliminates the possibility that this information triggers implicit associations or undermines egalitarian commitments (Bertrand and Mullainathan, 2004; Kawakami, Dovidio, and van Kamp, 2007; Washington and Kelly, 2016). It is a fine example of taking ecological control.

Unfortunately, it's not clear whether or how all possible gender cues could be removed from writing samples. In this case, our reader may choose to practice goal priming, for example, calling to mind her egalitarian commitments before sitting down to evaluate the philosophy papers. There is evidence to think that this exercise may facilitate goal pursuit even unconsciously (for example, see Moskowitz and Li, 2011). Another useful strategy involves implementation intentions. An implementation intention is an if-then plan one may rehearse, such as 'if I read a paragraph with a large number of pronouns, I will think "clear writing"' (since it has been suggested that women on average use more pronouns in their writing than men, Argamon et al., 2003). Practicing this intention will help automatize the consequent action – in this case, associating a style of syntax with good philosophical writing – in a way that avoids the negative association (Gollwitzer and Sheeran, 2006). Both of these strategies allow our reader to exercise ecological control. Her judgments will better align with her values without the need for reflective consideration of implicit bias.

Interestingly, there is evidence to think that an implementation intention of this kind may also do some work to alter the association between maleness and academic achievement. Another strategy in this vein is counter-stereotype exposure (Dasgupta and Greenwald, 2001). Reminders of successful women philosophers, because they are 'stereotype discordant', can similarly alter our reader's association, and thereby reduce its influence. Again, our reader manages her bias in an indirect way and respects the deeply rooted nature of implicit cognition.

Broader lessons about practical reasoning from the epistemology of implicit cognition

We contend that the four strategies presented previously, as examples of ecological control, are instances of rational action even if they are not instances of action caused

only by in-the-moment reflecting on reasons. Environmental scaffolding, goal priming, implementation intentions, and counter-stereotype exposure are all prescriptions for bringing behavior in line with one's values or reasons. Of course, they also shift one's focus outward – away from 'willpower' and toward the maintenance of one's cognitive ecology. The goal is to situate oneself in such a way that fast, automatic, and unconscious mechanisms can be harnessed toward better ends. Perhaps this is not surprising, given the efficient nature of implicit cognition, when compared with reflective, effortful deliberation. On the other hand, one might wonder whether these are prescriptions for better reasoning properly understood. Why should prescriptions for strengthening the reasoning capacities of individuals focus on institutional or cultural change?

To answer this question, two additional points about implicit cognition are worth making. First, implicit cognitive mechanisms are themselves a huge obstacle to overcoming their effects. That is, the mechanisms that drive implicit bias are the very same that hinder the conversation about how to ameliorate it. We are not only likely to be unaware of our implicit gender biases, we are liable to deny them, overestimating our own reflective powers in comparison to our peers. Bias Blind Spot and the Dunning-Kruger effect are two well-known examples of epistemic biases which can compound the problem of gender biases (Pronin, Lin, and Ross, 2002; Kruger and Dunning, 1999). The more influence these epistemic biases have, the less likely our reader is to exert ecological control over her gender biases. Second, it is worth taking into account where an implicit association between maleness and academic achievement comes from. Why is this negative association so prevalent, as opposed to any other association we would not endorse? The obvious answer is that the content of our cognitive biases is a product of the structural, cultural, and institutional biases surrounding us. Managing these injustices is the best way to facilitate egalitarian reasoning in creatures with limited introspective powers like ourselves. Thus, we argue that what at first glance may seem counter intuitive about the strategies we have presented is actually a feature that ought to be embraced.

4 Conclusion

In this chapter, we have reviewed three cases in which research in the social sciences has been taken to illuminate philosophical questions about practical reason. We think it's fair to say that empirical research is sometimes relevant and that whether it is or not and in what way will depend on the details of the subject; there is no general answer to the question about the philosophical relevance of the social sciences in this domain. Finally, when it comes to prescriptive recommendations about how to do better in ways that matter in philosophy, we think it is clear that empirical research about the obstacles to better behavior and effective strategies for overcoming them is invaluable.

References

Aharoni, E., Sinnott-Armstrong, W., and Kiehl, K. A. 2012. "Can Psychopathic Offenders Discern Moral Wrongs? A New Look at the Moral/Conventional Distinction." *Journal of Abnormal Psychology* 121 (2): 484.

Antony, L. 2012. "Different Voices or Perfect Storm: Why Are There So Few Women in Philosophy?" *Journal of Social Philosophy* 43 (3): 227–255.

Argamon, S., Koppel, M., Fine, J., and Shimoni, A. R. 2003. "Gender, Genre, and Writing Style in Formal Written Texts." *Text* 23 (3): 321–346.

Bertrand, M., and Mullainathan, S. 2004. "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market and Discrimination." *American Economic Review* 94 (4): 991–1013.

Bishop, M. A., and Trout, J. D. 2004. *Epistemology and the Psychology of Human Judgment*. Oxford: Oxford University Press.

Bjorklund, F., Haidt, J., and Murphy, S. 2000. "Moral Dumbfounding: When Intuition Finds No Reason." *Lund Psychological Reports* 2 (1): 29.

Blackburn, S. 1984. Spreading the Word: Groundings in the Philosophy of Language. *New York: Clarendon Press.*

Blair, R. J. R. 1995. "A Cognitive Developmental Approach to Morality: Investigating the Psychopath." *Cognition* 57 (1): 1–29.

Brownstein, M. 2015. "Implicit Bias." *The Stanford Encyclopedia of Philosophy* (Spring 2017 Edition), Edward N. Zalta (ed.). URL = <https://plato.stanford.edu/archives/spr2017/entries/implicit-bias/>.

Clark, A. 2007. "Soft Selves and Ecological Control." In D. Ross, D. Spurett, H. Kincaid, and G. L. Stephens (eds.), *Distributed Cognition and the Will* (pp. 101–122). Cambridge, MA: The MIT Press.

Covington, M. V. 2000. "Goal Theory, Motivation, and School Achievement: An Integrative Review." *Annual Review of Psychology* 51 (1): 171–200.

D'Arms, J., and Jacobson, D. 2000. "Sentiment and Value." *Ethics* 110 (4): 722–748.

Darley, J. M., and Batson, C. D. 1973. "From Jerusalem to Jericho: A Study of Situational and Dispositional Variables in Helping Behavior." *Journal of Personality and Social Psychology* 27 (1): 100.

Dasgupta, N., and Greenwald, A. G. 2001. "On the Malleability of Automatic Attitudes: Combating Automatic Prejudice with Images of Admired and Disliked Individuals." *Journal of Personality and Social Psychology* 81 (5): 800–814.

Doris, J. M. 2002. *Lack of Character: Personality and Moral Behavior*. Cambridge and New York: Cambridge University Press.

Doris, J. M. 2016. Talking to Our Selves: Reflection, Ignorance, and Agency. *Oxford: Oxford University Press.*

Ellis, S., Carette, B., Anseel, F., and Lievens, F. 2014. "Systematic Reflection Implications for Learning from Failures and Successes." *Current Directions in Psychological Science* 23 (1): 67–72.

Fischer, J. M., and Ravizza, M. 2000. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.

Gibbard, A. 1992. Wise Choices, Apt Feelings: A Theory of Normative Judgment. *Oxford: Oxford University Press.*

Gollwitzer, P. M., and Sheeran, P. 2006. "Implementation Intentions and Goal Achievement: A Meta-Analysis of Effects and Processes." *Advances in Experimental Social Psychology* 38: 69–119.

Greene, J. D. 2009. "Dual-Process Morality and the Personal/Impersonal Distinction: A Reply to McGuire, Langdon, Coltheart, and Mackenzie." *Journal of Experimental Social Psychology* 45 (3): 581–584.

Greene, J. D. 2013. Moral Tribes: Emotion, Reason, and the Gap Between Us and Them. *New York: Penguin.*

Haidt, J. 2001. "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment." *Psychological Review* 108 (4): 814–834.

Haidt, J., and Bjorklund, F. 2008. "Social Intuitionists Answer Six Questions about Morality." *Moral Psychology* 2: 181–217.

Halberstadt, J. B., and Wilson, T. 2008. "Reflections on Conscious Reflection: Mechanisms of Impairment by Reasons Analysis." In J. Adler and L. Rips (eds.), *Reasoning: Studies of Human Inference and Its Foundations* (pp. 548–565). Cambridge: Cambridge University Press.

Hare, R. D. 2003. *The Hare Psychopathy Checklist – Revised (PCL–R) Manual* (2nd ed.). Toronto, ON: Multi-Health Systems.

Holroyd, J. D. and Kelly, D. 2016. "Implicit Bias, Character and Control." In A. Masala and J. Webber (eds.), *From Personality to Virtue Essays on the Philosophy of Character*. Oxford: Oxford University Press.

Kahneman, D. 2011. *Thinking Fast and Slow*. New York: Farrar, Straus and Giroux.

Kawakami, K., Dovidio, J. F. and van Kamp, S. 2007. "The Impact of Näıve Theories Related to Strategies to Reduce Biases and Correction Processes on the Application of Stereotypes." *Group Processes and Intergroup Relation* 10: 139–156.

Kelly, D. (2011). Yuck!: The Nature and Moral Significance of Disgust. *Cambridge, MA: MIT Press.*

Kennett, J. 2006. "Do Psychopaths Really Threaten Moral Rationalism?" *Philosophical Explorations* 9 (1): 69–82.

Kruger, J., and Dunning, D. 1999. "Unskilled and Unaware of It: How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-Assessments." *Journal of Personality and Social Psychology* 77 (6): 1121–1134.

Latane, B., and Darley, J. M. 1970. *The Unresponsive Bystander: Why Doesn't He Help?* New York: Appleton-Century Crofts.

Lerner, J. S., Goldberg, J. H., and Tetlock, P. E. 1998. "Sober Second Thought: The Effects of Accountability, Anger, and Authoritarianism on Attributions of Responsibility." *Personality and Social Psychology Bulletin* 24 (6): 563–574.

Leslie, S. J., Cimpian, A., Meyer, M., and Freeland, E. 2015. "Expectations of Brilliance Underlie Gender Distributions Across Academic Disciplines." *Science* 347 (6219): 262–265.

Locke, E. A., and Latham, G. P. 2002. "Building a Practically Useful Theory of Goal Setting and Task Motivation: A 35-Year Odyssey." *American Psychologist* 57 (9): 705.

Morisano, D., Hirsh, J. B, Peterson, J. B., Pihl, R. O., and Shore, B. M. 2010. "Setting, Elaborating, and Reflecting on Personal Goals Improves Academic Performance." *Journal of Applied Psychology* 95 (2): 255.

Moskowitz, G. B. and Li, P. 2011. "Egalitarian Goals Trigger Stereotype Inhibition: A Proactive form of Stereotype Control." *Journal of Experimental Social Psychology* 47 (1): 103–116.

Nichols, S. 2004. Sentimental Rules: On the Natural Foundations of Moral Judgment*. Vol. 13. Oxford: Oxford University Press.*

———. 2010. "How Psychopaths Threaten Moral Rationalism." *Moral Psychology: Historical and Contemporary Readings* 73.

Prinz, J. 2006. "The Emotional Basis of Moral Judgments." *Philosophical Explorations* 9 (1) (March): 29–43.

Prinz, J. 2007. *The Emotional Construction of Morals*. Oxford: Oxford University Press.

Pronin, E., Lin, D. Y., and Ross, L. 2002. "The Bias Blind Spot: Perceptions of Bias in Self Versus Others." *Personality and Social Psychology Bulletin* 28: 369–381.

Roskies, A. 2006. "Neuroscientific Challenges to Free Will and Responsibility." *Trends in Cognitive Sciences* 10(9): 419–423.

Saul, J. M. 2017. "Why So Few Women in Value Journals? How Could We Find Out?" *Public Affairs Quarterly* (April 1): 125–141.

Schippers, M. C., Scheepers, A. W. A., and Peterson, J. B. 2015. "A Scalable Goal-Setting Intervention Closes Both the Gender and Ethnic Minority Achievement Gap." Palgrave Communications 1. www.palgrave-journals.com/articles/palcomms201514.

Schnall, S., Haidt, J., Clore, G. L., and Jordan, A. H. 2008. "Disgust as Embodied Moral Judgment." *Personality and Social Psychology Bulletin* 34 (8): 1096–1109.

Seidel, A., and Prinz, J. 2013. "Sound Morality: Irritating and Icky Noises Amplify Judgments in Divergent Moral Domains." *Cognition* 127 (1): 1–5.

Sommers, T. 2005. "Interview with Jonathan Haidt." *The Believer*. <www.believermag.com/issues/200508/?read=interview_haidt>.

Stein, E. 1996. Without Good Reason: The Rationality Debate in Philosophy and Cognitive Science. *Clarendon Press.*

Tiberius, V. 2016. "Does the New Wave in Moral Psychology Sink Kant?" In Kelly James Clark (ed.), *The Blackwell Handbook on Naturalism* (pp. 336–350). West Sussex, UK: Wiley Blackwell.

Tiberius, V. 2015. Moral Psychology: A Contemporary Introduction. *New York: Routledge.*

Vargas, M. (2013). Building Better Beings: A Theory of Moral Responsibility. *Oxford: Oxford University Press.*

Washington, N., and Kelly, D. 2016. "Who's Responsible for This? Moral Responsibility, Externalism, and Knowledge about Implicit Bias." In M. Brownstein and J. Saul (eds.), *Implicit Bias and Philosophy.* New York: Oxford University Press.

Wheatley, T., and Haidt, J. 2005. "Hypnotic Disgust Makes Moral Judgments More Severe." *Psychological Science* 16 (10): 780–784.

Wilson, T. D., andNisbett, R. E. 1978. "The Accuracy of Verbal Reports about the Effects of Stimuli on Evaluations and Behavior." *Social Psychology; Social Psychology.* http://psycnet.apa.org/psycinfo/1980-24471-001.

Wilson, T. D., and Kraft, D. 1993. "Why Do I Love Thee? Effects of Repeated Introspections about a Dating Relationship on Attitudes Toward the Relationship." *Personality and Social Psychology Bulletin* 19 (4): 409–418.

Wilson, T. D., Kraft, D., and Dunn, D. S. 1989. "The Disruptive Effects of Explaining Attitudes: The Moderating Effect of Knowledge about the Attitude Object." *Journal of Experimental Social Psychology* 25 (5): 379–400.

Wilson, T. D., Lisle, D. J., Schooler, J. W., Hodges, S. D., Klaaren, K. J., and LaFleur, S. J. 1993. "Introspecting about Reasons Can Reduce Post-Choice Satisfaction." *Personality and Social Psychology Bulletin* 19: 331–331.

Part 4: The philosophy of practical reason as the theory of practical normativity

Section 1: The nature and grounds of normative practical reasons

20. Objectivism about reasons[(2)]

[(2)] Prefatory note from Ruth Chang: Two weeks before his sudden death on January 1, 2017, Derek Parfit contacted me about his contribution to this volume. He said he planned to discuss two topics: his theory of normative practical reasons and his metanormative view about the nature of normative facts and claims. He also said that he expected to quote extensively from his already-published *On What Matters*, volumes 1–3 (Oxford: Oxford University Press, 2011 (vols. 1 &2), 2017 (vol. 3).

Derek Parfit (edited by Ruth Chang)

I have tried to honor Parfit's intentions by carving two papers from *On What Matters*, one on each of the topics he intended to write about, with a minimum of editorial intervention. This chapter gives Parfit's main defense of the idea that all normative practical reasons are 'value-based' and not 'desire-based'. A companion chapter, 'Parfit on Normative Nonnaturalism' describes and argues for Parfit's metanormative view, a view he was increasingly excited by toward the end of his life because he thought that it was a truth on which seemingly competing metaethical views could – and he hoped would – eventually converge. Both chapters are condensed and slightly modified reprints of material from key chapters in *On What Matters.*

Peter Momtchiloff of Oxford University Press kindly agreed to allow Parfit's material to be re-purposed for this volume on the condition that the authorship of the paper indicated that the material had been edited by me. Needless to say, the work here is entirely Parfit's. We hope that this condensed version of Parfit's arguments will lead readers to investigate the original text.

* * *

At the heart of Parfit's magisterial three-volume work, *On What Matters*, is the striking conclusion that three seemingly fundamentally opposed theories of morality – consequentialism, deontology, and Scanlonian contractualism – are in fact three approaches to a single, unified, and correct theory about how we should live, what Parfit calls 'the Triple Theory'. According to the Triple Theory, an act is wrong "just when such acts are disallowed by some principle that is optimific, uniquely universally willable, and not reasonably rejectable" (vol. 1, 413). We should all live our lives, Parfit urges, so as to avoid such wrongdoing.

Parfit's argument crucially depends on a deeper, more general question. What reasons do we have to do or to want things? In this paper, Parfit lays out his Objectivist theory of reasons and gives what he regarded as the main argument against the opposing Subjectivist theory of reasons. Readers interested in further arguments against Subjectivism as well as a diagnosis of why philosophers have been attracted to the view, should read Parfit's original text, chapters 1–7 (vol. 1) and 49 (vol. 3) of On What Matters.

Objectivism about reasons is of general interest to understanding practical reason not only for the role it plays in Parfit's argument for the Triple Theory of morality but also more generally because a) it presumes a dichotomy between reasons as being either value-based or desire-based, while some have argued that there can be will-based reasons, and b) it presumes that rationality is to be explained in terms of reasons and not the other way around, while some have argued that we should understand what reasons we have in terms of what it is rational for us to do. Both assumptions are challenged by philosophers in this volume. While reading this chapter, students might also be aware of two distinct questions: i) What kinds of considerations can *be* normative reasons? And ii) In virtue of what is that consideration a reason?

What follows is Parfit's case for the idea that the normative reasons we have to do things are "given" by facts that make actions and the objects of our desires valuable in some way and not our desires themselves.

1 Two kinds of theory of practical reasons

> There are two main kinds of view about what I shall call practical reasons. According to one group of views, there are certain facts that give us reasons both to have certain desires and aims, and to do whatever might achieve these aims. These reasons are given by facts about the objects of these desires or aims, or what we might want or try to achieve. We can therefore call such reasons object-given. If we believe that all practical reasons are of this kind, we are Objectivists about Reasons, who accept or assume some objective theory.

Object-given reasons are provided by the facts that make certain outcomes worth producing or preventing, or make certain things worth doing for their own sake. In most cases, these reason-giving facts also make these outcomes or acts good or bad for particular people, or impersonally good or bad. So we can also call these objective reasons and theories value-based.

According to another group of theories, our reasons for acting are all provided by, or depend upon, certain facts about what would fulfil or achieve our present desires or aims. Some of these theories appeal to our actual present desires or aims. Others appeal to the desires or aims that we would now have, or to the choices that we would now make, if we had carefully considered all of the relevant facts. Since these are all facts about us, we can call these reasons subject-given. If we believe that all practical reasons are of this kind, we are Subjectivists about Reasons, who accept some subjective theory.

These two kinds of theory are very different. According to Objectivists, though many reasons for acting can be claimed to be given by the fact that some act would achieve one of our aims, these reasons derive their force from the facts that give us reasons to have these aims. These are the facts that make these aims relevantly good, or worth achieving.

According to Subjectivists, we have no such reasons to have our aims. Some Subjectivists even claim that it is we who, with our desires or choices, make things good. While defending such a view, for example, Korsgaard writes:

> most things are good because of the interest human beings have in them ... Objectivism reverses this relation ... Instead of saying that what we are

interested in is therefore good, the objectivist says that the goodness is in the object, and we ought therefore to be interested in it.

Such goodness would give us reasons in the way the sun gives light, 'because it's out there, shining down'. If Subjectivism is true, we must make our choices in the dark.

(vol. 1, 45–46)

The same facts can give us reasons both to want something to happen and to try to make it happen by acting in some way. That is why I call both kinds of reason practical. Though these two kinds of reason are very closely related, there is a striking difference between the ways in which we can respond to them. When we are aware of facts that give us reasons to act in some way, we can often respond to these reasons by acting in this way. This response is voluntary in the sense that, if we had wanted not to act in this way, we could have chosen not to do so. But when we are aware of facts that give us strong reasons to have some desire, our response to these reasons is seldom voluntary. It is seldom true that, if we had wanted not to have such desires, we could have chosen not to have them. We could seldom choose, for example, whether we want to stay alive, or to avoid great pain. If some whimsical despot threatens to kill me unless, one minute from now, I want to be killed, I could not choose to have this desire.

(vol. 1, 47)

Our reasons to have some desire are provided, I have claimed, by facts about this desire's object, or the event that we want. Such reasons I am calling object-given. Many people assume that we can also have state-given reasons to have some desire. Such reasons would be provided by certain facts, not about some desire's object, but about our state of having this desire. We would have such reasons when our having some desire would be in some way good, either as an end or as a means.

In this view, we can have at least four kinds of reason to have some desire, which can be described as follows:

	telic and intrinsic	instrumental
object-given	The event that we want would be in itself good, or worth achieving	This event would have good effects
state-given	Our wanting this event would be in itself good	Our wanting this event would have good effects

We might have reasons of all these kinds to have the same desire. If you are in pain, for example, I might have all these reasons to want your pain to end. What I want

would be in itself good, and it might also have the good effect of allowing you to enjoy life again. My wanting your pain to end might be in itself good, and this desire might also have good effects, such as your being comforted by my sympathy.

(vol. 1, 50)

> [State-given] ... reasons would not, I believe, have any importance. When it would be better if we were in some state, we would have reasons to want to be in this state. If we could cause ourselves to be in this state, we would have reasons to act in this way. It is not worth claiming that, as well as having reasons to *want* to be and to *cause* ourselves to be in this state, we would also have reasons to *be* in this state. Suppose for example that I would be healthier and happier if I weighed less, owned a bicycle, knew how to dance, and had some friends. These facts would give me reasons to want and to try to make myself lose weight, to buy a bicycle, to learn how to dance, and to make some friends. It is not worth claiming that, as well giving me reasons to act in these ways, these facts would give me reasons to weigh less, to own a bicycle, to know how to dance, and to have some friends. Such reasons would make no difference.

Suppose next that, though it would be better if we were in a certain state, we could not possibly cause ourselves to be in this state. We would then have reasons to wish that we were in this better state. I might have reasons, for example, to wish that I were ten inches taller, twenty years younger, and could run faster than a cheetah. We needn't claim that I would also have reasons to *be* ten inches taller, to *be* twenty years younger, and to *be able* to run faster than a cheetah. And such claims may be clearly false. Reasons are things to which at least some people might be able to respond, and no one could respond to a reason to be twenty years younger.

Similar claims apply to our beliefs and desires. When it would be better for us if we had some belief or desire, we have object-given reasons to want to have this belief or desire, and to cause ourselves to have it, if we can. It is not worth claiming that we also have state-given reasons to *have* this belief or desire.

(vol. 1, 51)

2 Subjective theories of reasons

> Subjective theories appeal to facts about our present desires, aims, and choices." (58). "... [O]ur desires are *telic* when we want some event as an end, or for its own sake, and *instrumental* when we want some event as a means to some end. Our *aims* are often the telic desires that we have decided to try to fulfill.
>
> (vol. 1, 58–59)
>
> "According to the Informed Desire Theory: We have most reason to do whatever would best fulfil the telic desires or aims that we would now have if we knew all of the relevant facts.

Any fact counts as *relevant*, some writers claim, if our knowledge of this fact would affect our desires. But this criterion is too wide. As Gibbard remarks, if we knew and vividly imagined the full facts about what is going on in the innards of our fellow-diners, we might lose our desire to eat. And if we learnt certain facts about man's inhumanity to man, we might become so depressed that we would lose our desire to live. The Informed Desire Theory would then implausibly imply that, even though we actually want to eat and to stay alive, we have no reason to fulfil these desires. To avoid such implications, some Subjectivists claim that, for some fact to count as *relevant*, it is not enough that our knowledge of this fact would affect our desires. In such views, when we are choosing between several possible acts, what are relevant are only facts about these acts and their possible outcomes.

The Informed Desire Theory needs another revision. It is sometimes true that, if we were fully informed, that would change our situation in some way that altered both our desires and what we had reasons to do. If Subjectivists claim that our reasons are provided, not by our actual desires, but by our hypothetical informed desires, these people may be led in such cases to implausible conclusions. Suppose, for example, that we want to learn certain important facts. If we knew these facts, we would lose this desire. But that should not be taken to imply that we have no reason to act on this desire, by trying to learn these facts. Some Subjectivists therefore claim that we should try to fulfil the desires that, if we were fully informed, we would want ourselves to have in our actual uninformed state.

Some other Subjectivists appeal, not to what would best fulfil or achieve our desires or aims, but to the choices or decisions that we would make after carefully considering the facts. These people also make claims about how it would be rational for us to make

such decisions. According to what we can call the Deliberative Theory: We have most reason to do whatever, after fully informed and rational deliberation, we would choose to do.

This form of Subjectivism can be easily confused with Objectivism, since such theories can be stated in deceptively similar ways. Subjectivists and Objectivists might both claim that

> (A) what we have most reason to do, or decisive reasons to do, is the same as what, if we were fully informed and rational, we would choose to do.

But this claim is ambiguous. Subjectivists and Objectivists may both claim that, when we are trying to make some important decision, we ought to deliberate in certain ways. We ought to try to imagine fully the important effects of our different possible acts, to avoid wishful thinking, to assess probabilities correctly, and to follow certain other procedural rules. If we deliberate in these ways, we are *procedurally* rational.

Objectivists make further claims about the desires and aims that we would have, and the choices that we would make, if we were also *substantively* rational. These claims are *substantive* in the sense that they are not about *how* we make our choices, but about *what* we choose. There are various telic desires and aims, Objectivists believe, that we all have strong and often decisive object-given reasons to have. To be fully substantively rational, we must respond to these reasons by having these desires and aims, and trying to fulfil or achieve them if we can. Deliberative Subjectivists make no such claims. These people deny that we have such object-given reasons, and they appeal to claims that are only about procedural rationality.

Though these two groups of people might both accept (A), they would explain (A) in different ways.

According to these Subjectivists, when it is true that

> (B) if we were fully informed and procedurally rational, we would choose to act in some way,

this fact makes it true that

> (C) we have decisive reasons to act in this way.

Objectivists claim instead that, when it is true that

> (D) we have decisive reasons to act in some way,

this fact makes it true that

> (E) if we were fully informed and both procedurally and substantively rational, we would choose to act in this way.

To illustrate these claims, we can suppose that, unless I stop smoking, I shall die much younger, losing many years of happy life. According to all plausible objective theories, this fact gives me a decisive reason to want and to try to stop smoking. If I were fully informed and substantively rational, that is what I would choose to do. What we ought rationally to choose, Objectivists believe, depends on what we have such reasons or apparent reasons to want and to do.

Suppose next that, after fully informed and procedurally rational deliberation – or what we can now call *ideal* deliberation – I would choose to stop smoking. Deliberative Subjectivists would then agree that I have a decisive reason to stop smoking. In this view, however, the inference runs the other way. Instead of claiming that what we ought to choose depends on our reasons, these Subjectivists claim that our reasons depend on what, after such deliberation, we would choose. If I have decisive reasons to stop smoking, that is *because* I would choose to act in this way.

As this example shows, these theories are very different. These Objectivists appeal to normative claims about what, after ideal deliberation, we have *reasons* to choose, and *ought rationally* to choose. These Subjectivists appeal to psychological claims about what, after such deliberation, we *would in fact* choose.

(vol. 1, 60–63)

3 Why subjective theories are mistaken: the agony argument

> Subjective theories can have implausible implications. Suppose that[:] ...
>
> *Case:* I know that some future event would cause me to have some period of agony. Even after ideal deliberation, I have no desire to avoid this agony. Nor do I have any other desire or aim whose fulfilment would be prevented either by this agony, or by my having no desire to avoid this agony.

Since I have no such desire or aim, all subjective theories imply that I have no reason to want to avoid this agony, and no reason to try to avoid it, if I can.

[... [*Case*] might be claimed to be impossible, because my state of mind would not be agony unless I had a strong desire not to be in this state. But this objection overlooks the difference between our attitudes to present and future agony. Though I know that, when I am later in agony, I shall have a strong desire not to be in this state, I might have no desire now to avoid this future agony.

It might next be claimed that my predictable future desire not to be in agony gives me a desire-based reason now to want to avoid this future agony. But this claim cannot be made by those who accept subjective theories of the kind that we are considering. These people do not claim, and given their other assumptions they could not claim, that facts about our future desires give us reasons.

Some other theories make that claim. A value-based objective theory about reasons might be combined with a desire-based subjective theory about well-being. On such a view, even if we don't now care about our future well-being, we have reasons to care, and we ought to care. These reasons are value-based in the sense that they are provided by the facts that would make various future events good or bad for us. But if our future well-being would in part consist, as this view claims, in the fulfilment of some of our future desires, these value-based reasons would be reasons to act in ways that would cause these future desires to be fulfilled. It might be similarly claimed that we have value-based reasons to fulfil other people's desires, because such acts would promote the well-being of these other people. Though these theories claim that we have reasons to fulfil these desires, these value-based objective theories about reasons are very different from the desire-based subjective theories that we are now considering.

We can also imagine a temporally neutral desire-based theory. On this view, what we have most reason to do, at any time, is whatever would best fulfil all of our desires throughout our life, whether or not these acts would be good for us. According to a

similar, personally neutral theory, what we have most reason to do is whatever would best fulfil everyone's desires, whether or not these acts would be good for anyone. These imagined theories are also very different from the subjective theories that we are now considering.

According to these theories, it is only certain facts about our own present desires, aims, or choices that give us reasons, or on which our reasons depend. We are supposing that, in *Case*, I have carefully considered all of the relevant facts about my possible future period of agony. Since I have no present desire or aim whose fulfilment would be prevented either by this agony, or by my having no desire to avoid this agony, all subjective theories imply that I have no reason to want to avoid this agony. Similar claims apply to my acts. Even if I could easily avoid this agony – perhaps by moving my hand away from the flames of some approaching fire – I have no reason to act in this way. Such a reason would have to be provided by some relevant present desire, and I have no such desire.

> (vol. 1, 73–75)
>
> Subjectivists agree that it would make sense to claim that I have a reason to want and to try to avoid this future agony. But these people's theories imply that, since I have no relevant present desire, I have no such reason. No fact counts in favour of my wanting and trying to avoid this agony. Similar claims apply to other such cases. According to these Subjectivists, when we have no relevant present desires, we would have no reason to want to avoid some period of future agony.

We can now argue:

> We all have a reason to want to avoid, and to try to avoid, all future agony.
>
> Subjectivism implies that we have no such reason.

Therefore

> Subjectivism is false.
>
> We can call this the *Agony Argument*.

Some Subjectivists might claim that we can ignore this argument, because my example is purely imaginary. Every actual person, they might say, wants to avoid all future agony.

This reply would fail. First, we are asking whether subjective theories imply that we all have a reason to want to avoid all future agony. To support the claim that we all have such a reason, it is not enough to claim that everyone has this desire. These Subjectivists would also have to claim that, when we have some desire, this fact gives us a reason to have it. As we shall see, that is an indefensible claim.

Second, it seems likely that some actual people do not want to avoid all future agony. Many people care very little about pain in the further future. Of those who have believed that sinners would be punished with agony in Hell, many tried to stop sinning only when they became ill, and Hell seemed near. And when some people are very depressed, they cease to care about their future well-being.

Third, even if there were no such actual cases, normative theories ought to have acceptable implications in merely imagined cases, when it is clear enough what such cases would involve. Subjectivists make claims about which facts give us reasons. These claims cannot be true in the actual world unless they would also have been true in possible worlds in which there were people who were like us, except that these people did not want to avoid all future agony, or their desires differed from ours in certain other ways. So we can fairly test subjective theories by considering such cases.

Subjectivists might reply that, even in such possible worlds, there would be some telic desires that everyone must have, because without these desires these people could not even be rational agents, who can act for reasons. To be such agents, Williams suggests, we must have 'a desire not to fail through error', and some 'modest amount of prudence'. But such claims are irrelevant here. We could be agents who act for reasons without wanting to avoid all future agony.

Subjectivists might next claim that, if some theory has acceptable implications in all or most actual cases, this fact may give us sufficient reasons to accept this theory. We might justifiably accept such a theory even if there are some unusual or imagined cases in which this theory's implications seem to be mistaken. Many theories of many kinds can be plausibly defended in this way. For such a defence to succeed, however, we must be able to claim that there are no other, competing theories which have more acceptable implications. And Subjectivists cannot make that claim. When subjective theories are applied to actual people, these theories often have plausible implications. But that is because most actual people often have desires that they have object-given reasons to have, because they want things that are in some way good, or worth achieving. In many such cases, subjective theories have the same implications as the best objective theories. In trying to decide which theories are best, we must consider cases in which these two kinds of theory disagree. That is how, for similar reasons, we must decide between different scientific theories. Such disagreements take their clearest form in some unusual actual cases and some imaginary cases. So Subjectivists cannot claim that we can ignore these cases, or that we can give less weight to them. On the contrary, these are precisely the cases that we have most reason to consider. In their claims about such cases, subjective theories are, I am arguing, much less plausible than the best objective theories. And if these objective theories are more plausible whenever these two kinds of theory disagree, these objective theories are clearly better.

There is another possible reply. Deliberative Subjectivists appeal to what we would want and choose after some process of informed and rational deliberation. These people might argue:

(A) We all have reasons to have those desires that would be had by anyone who was fully rational.

(B) Anyone who was fully rational would want to avoid all future agony.

Therefore

We all have a reason to want to avoid all future agony. As I have said, however, such claims are ambiguous. Objectivists could accept (B), because these people make claims about substantive rationality. According to objective theories, we all have decisive reasons to have certain desires, and to be substantively rational we must have these desires. These reasons are given by the intrinsic features of what we might want, or might want to avoid. We have such a decisive object given reason to want to avoid all future agony. If we did not have this desire, we would not be fully substantively rational, because we would be failing to respond to this reason.

Subjectivists cannot, however, make such claims. On subjective theories, we have no such object-given reasons, not even reasons to want to avoid future agony. Deliberative Subjectivists appeal to what we would want after deliberation that was merely procedurally rational. On these theories, if we have certain telic desires or aims, we may be rationally required to want, and to do, what would best fulfil or achieve these desires or aims. But, except perhaps for the few desires without which we could not even be agents, there are no telic desires or aims that we are rationally required to have. We can be procedurally rational whatever else we care about, or want to achieve. As one Subjectivist, [John] Rawls, writes:

knowing that people are rational, we do not know the ends they will pursue, only that they will pursue them intelligently.

So Subjectivists cannot claim that anyone who is fully rational would want to avoid all future agony.

It might be objected that, in making these remarks, I have underestimated what Subjectivists can achieve by appealing to claims about procedural rationality. [Michael] Smith, for example, claims that

(C) we are rationally required not to have desires or preferences that draw some arbitrary distinction.

By appealing to this 'minimal principle', Smith writes, Subjectivists can explain the irrationality of many desires and preferences, such as the preferences of my imagined man who cares about what will happen to him except on any future Tuesday. This man's preferences are irrational, Smith claims, because they draw an arbitrary distinc-

tion. It would be similarly arbitrary, Subjectivists might claim, not to want to avoid all future agony.

Subjectivists cannot, however, make such claims. Our preferences draw arbitrary distinctions when, and because, what we prefer is in no way preferable. It is arbitrary to prefer one of two things if there are no facts about these things that give us any reason to have this preference. My imagined man would prefer to have one of two similar ordeals if, and because, this ordeal would be on a future Tuesday. To explain why this preference is arbitrary, we must claim that

> (1) if some ordeal would be on a future Tuesday, this fact does not give us any reason to care about it less.

Unlike my imagined man, most of us would always prefer to have one of two ordeals if, and because, this ordeal would be less painful. To explain why this preference is not arbitrary, we must claim that

> (2) if some ordeal would be less painful, this fact does give us a reason to care about it less.

(1) and (2) are claims about object-given reasons. Since Subjectivists deny that we have such reasons, these people cannot appeal to such claims, or to the 'minimal principle' that Smith states with (C).

Smith also claims that

> (D) we can be rationally required to have some desire when, and because, our having this desire would make our set of desires more coherent and unified.

To illustrate this requirement, Smith supposes that we want to help only some of the people whom we know to be in desperate need. Our desires would be more coherent, and would 'make more sense', Smith claims, if we wanted to help all of these people. But this claim assumes that

> (3) whenever someone is in desperate need, this fact gives us a reason to want to help this person.

If such facts did not give us such reasons, our desires would not be less coherent, or make less sense, if we wanted to help only some of these people. And (3) is another claim about object-given reasons, to which Subjectivists cannot appeal.

Consider next Smith's claim that we can be rationally required to have a more unified set of desires. Mere unity is not a merit. Our desires would be more unified if we were monomaniacs, who cared about only one thing. But if you cared about truth, beauty, and the future of humanity and I cared only about my stamp collection, your

less unified set of desires would not be, as Smith's claim seems to imply, less rational than mine. Smith might reply that my set of desires would be more impressively unified if I had several coherent desires. But if I also wanted to collect match-boxes, drawing pins, ticket stubs, and plastic cups, your less unified set of desires would still be more rational than mine. And this appeal to coherence would again assume that we have object-given reasons to have our desires. Subjectivists deny that we have such reasons.

(vol.1, 75–81)

> Smith ... suggest[s] another reply. If we were fully procedurally rational ... and we had a maximally coherent psychology, we would want to exercise our capacities to have rational beliefs and to fulfil our other desires, and we would therefore want to avoid anything that would interfere with our exercise of these capacities. Since our being in agony would involve such interference, we would want to avoid all future agony ...

Subjectivists, I agreed, could claim that we have instrumental reasons to have some desire when, and because, that would help us to fulfil one of our other present desires. My objection was that, on Subjectivist theories, the nature of agony give us no reason to have a *telic* desire, wanting to avoid future agony, not as a means of fulfilling some other present desire, but for its own sake, or as an end. I asked: 'Why can't we have such a reason?' Smith's ... argument could not answer this question. Since Smith doesn't even claim that, if we were rational, we would *have* such telic desires to avoid future agony, his claims could not imply that we have any *reason* to have such desires.

(vol. 3, 252)

> [Smith's view implies something hard to believe.] Suppose ... that we need some kind of surgery, during which we could either be anaesthetized and unconscious or be in agony. In Smith's view, we would have no reason to prefer to be unconscious rather than in agony, since both states would equally interfere with our exercise of our rational capacities.

(vol. 3, 252)

> When Smith heard me say some years ago that the nature of agony gives us a reason to want to avoid future agony, Smith remarked, if I remember right, that my claim was a paradigm of the kind of claim that good philosophers ought not to make. This remark puzzled me. But that was before I realized that several very good philosophers do not use the concept of a purely normative reason. If we don't use this concept, we cannot believe that the nature of agony gives us a reason to want to avoid future agony. Smith seems not to use this concept.

(vol. 3, 253)

"For all these reasons, Subjectivists cannot claim that, if we were procedurally rational, we would want to avoid all future agony.

(vol. 1, 80)

Since Subjectivists cannot defend this claim, my earlier conclusion stands. Subjectivists must claim that, in *Case*, I would have no reason to want to avoid my future period of agony. As I have said, we can argue:

We all have a reason to want to avoid, and to try to avoid, all future agony.

Subjectivism implies that we have no such reason.

Therefore
Subjectivism is false.

Some Subjectivists might now bite the bullet, by denying that we have this reason. In *Case*, these people might say, though the approaching flames threaten to cause me excruciating pain, this fact does not count in favour of my wanting and trying to move my hand away. But that is hard to believe.

We can next remember why Subjectivism has these implications. Since Subjectivists deny that we have object-given reasons, they must agree that, on their view,

(E) the nature of agony gives us no reason to want to avoid being in agony.

We can argue:
The nature of agony does give us such a reason.
Therefore
Subjectivism is false.
These arguments are, I believe, decisive.

Subjectivists might protest that, in denying (E), we are not arguing against their view, but are merely rejecting this view. If that is so, our claim could instead be that everyone ought to reject this view, since (E) is a very implausible belief. Subjectivists are not Nihilists, who deny that we have any reasons. These people believe that we have reasons for acting. If we can have some reasons, nothing is clearer than the truth that, in the reason-implying sense, it is bad to be in agony. It can be hard to remember accurately what it was like to have sensations that were intensely painful. Some of the awfulness disappears. But we can remember such experiences well enough. According to Subjectivists, what we remember gives us no reason to want to avoid having such intense pain again. If we ask 'Why not?', Subjectivist have, I believe, no good reply.

(vol.1, 81–82)

21. How to be a subjectivist[1]

[1] Thanks to Ruth Chang and Kurt Sylvan for very helpful comments on earlier versions of this chapter.

David Sobel

The *Euthyphro* poses perhaps the most basic and profound question about why things are valuable: Are things valuable because they are valued, or are things appropriately valued because they are valuable? Do our attitudes, as Hume says, gild and stain the world with value, or are there already established normative facts prior to our stance towards the world that determine what we ought to value? Subjectivists are the folks who think the former is the case.

The lesson of the *Euthyphro* was that the fully subjectivist direction of explanation looks misguided in some normative realms. The sensible subjectivist that I want to discuss here accepts that our attitudes are not a plausible ground of truth for all kinds of normativity, most obviously morality, for just the reasons Socrates outlined. But our subjectivist is more optimistic about the prospects of the attitudes grounding normativity in other, more individualized, contexts. Two contexts that have been especially tempting areas for subjectivist theorists are well-being and reasons for action. Our subjectivist will claim that in these contexts it is the agent's own attitudes that determine what is good for her or what she has reason to do.

After very briefly discussing what the more plausible versions of subjectivism might look like, and mentioning and contrasting a few neighboring views, we will get down to describing how a subjectivist might try to defend her view.

What is valuing?

Our subjectivist thinks that attitudes ground some kinds of normativity. But which attitudes in particular should she point to as the most plausible candidates? Let's call the relevant attitude valuing. What is it to value something in this sense? Clearly it is to take some sort of psychological attitude towards an object. Plausibly, just believing something is valuable is one way of valuing an option. But such valuing states, to my mind, are less plausible as accounts of what grounds value.[1] Typical beliefs that something is valuable seem to presuppose some other ground of value that could make such beliefs true. Such beliefs, like beliefs generally, appear to us to be attempts to get something independent of our belief correct. It is not very tempting to my mind to think that things are valuable just because we think they are. If we seek a conception of valuing that is more tempting as a ground of value, I think our subjectivist should look to conative, not truth-assessable, favorings such as liking, loving, wanting, or desiring. It sounds more plausible to say that chocolate ice cream is better for me, other things equal, than vanilla because I like it better, rather than because I believe it is better for me. So I think subjectivists should be searching for a conative attitude to play the value grounding role.

Often subjectivists maintain that the authoritative conative favoring attitudes are those that are factually informed, or at least not factually misinformed, about their object. My desire for that drink over there, which I reasonably believe is a gin and tonic but is in fact a petrol and tonic, seems to mislead me about what I have good reason to do. Thus many subjectivists are tempted to idealize the desires granted normative authority.[2]

A common subjectivist method of idealizing is to focus on the relevant valuing attitudes the agent would have after she was fully appreciative of what the option in question, or all options, would be like for her, in addition to requiring that the agent be fully factually informed.[3] But to remain a subjectivist view, this idealization needs

[1] But see Dale Dorsey, "Subjectivism Without Desire," *Philosophical Review* 121(3) (2012): 407–42; Ruth Chang, "Voluntarist Reasons and the Sources of Normativity," in *Reasons for Action*, eds. Sobel and Wall (Cambridge: Cambridge University Press, 2009), 243–271; and Sharon Street, "How to Be a Relativist About Normativity," manuscript.

[2] "Desire" here is a term of art that is meant as a general term for conative favorings, not a specific version of such favorings.

[3] The great line of early developers (I do not claim, in each case, champions) of subjectivist views assumed that informed desires were the normatively relevant ones. Hume, Mill, and Sidgwick, among others, all factually idealized the desires thought to be normatively relevant to our reasons or well-being. See David Hume, *A Treatise of Human Nature*, ed. Shelby-Bigge (Oxford: Oxford University Press, 1967),

to be procedural rather than substantive. This notion of procedural idealization is not trivial to explicate and perhaps not fully understood.[4]

A straightforwardly substantive idealization would require that one desires those things that are worthy of desire regardless of our motivational states. And a straightforward procedural idealization would require that one's preferences be, for example, internally coherent, or based on accurate forecasts of what the option would in fact be like.[5] A proper procedural idealization should not directly rule out the possibility of any particular option being the object of the idealized valuing attitude.

The intuitive idea is that a procedural account of idealization will not presuppose, and build this presupposition into the content of the idealization, that certain goods are worthier of the idealized valuing attitude than others. But saying only so much would have it that a view that privileged one's desires on Tuesday, while wearing a hat, or when on heroin would still count as relevantly procedural. Such views continue to look procedural in the sense that they do not presuppose that certain objects are worthier of the favoring attitudes than others. Yet these sorts of idealizations seem arbitrary in that they seem to not plausibly help capture the agent's own evaluative point of view.

The sort of procedural idealization our subjectivist has in mind aspires to plausibly explicate and reveal the agent's genuine concerns, not randomly privilege a class of

460; J.S. Mill, *Utilitarianism* (Indianapolis: Hackett Publishing, 2002), Chapter 2; Henry Sidgwick, *The Methods of Ethics*, 7th ed. (Indianapolis: Hackett, 1981), 111–12. See also Richard Brandt, *A Theory of the Good and the Right* (New York: Prometheus, 1979), 10, 113, 329; John Harsanyi, "Morality and the Theory of Rational Behavior," in *Utilitarianism and Beyond*, eds. Amartya Sen and Bernard Williams (Cambridge: Cambridge University Press, 1973), 55; John Rawls, *A Theory of Justice* (Cambridge, MA: Belknap, 1971), 407–24; Richard Hare, *Moral Thinking* (Oxford: Clarendon, 1981), 101–5 and 214–16. See also Douglas Senor, N. Fotion, and Richard Hare, eds., *Hare and Critics* (Oxford: Clarendon, 1990), 217–18; Peter Railton, "Facts and Values," *Philosophical Topics* 14 (1986): 5–29; David Gauthier, *Morals by Agreement* (Oxford: Clarendon, 1986), chap. 2; James Griffin, *Well-Being* (Oxford: Oxford University Press, 1986), 11–17; Shelly Kagan, *The Limits of Morality* (Oxford: Oxford University Press, 1989), 283–91. Comparable accounts of practical reasons have been influentially championed by (albeit sometimes in a Kantian rather than Humean spirit) Bernard Williams, "Internal and External Reasons," in his *Moral Luck* (Oxford: Oxford University Press, 1981), 101–13; Stephen Darwall, *Impartial Reason* (Ithaca, NY: Cornell University Press, 1983), pt. 2; David Lewis, "Dispositional Theories of Value," suppl. ser., *Proceedings of the Aristotelian Society* 63 (1989): 113–37; Michael Smith, *The Moral Problem* (Oxford: Blackwell, 1994); Julia Markovits, *Moral Reason* (Oxford: Oxford University Press, 2014).

[4] See, for example, Brad Hooker and Bart Streumer, "Procedural and Substantive Practical Rationality," in the *Oxford Handbook of Practical Rationality*, eds. Alfred Mele and Piers Rawling (Oxford: Oxford University Press, 2004).

[5] The right of subjectivists to appeal to idealized desires has increasingly been challenged. I reply to such worries in my "Subjectivism and Idealization," in my *From Valuing to Value* (Oxford: Oxford University Press, 2017). For the challenge see David Enoch, "Why Idealize?" *Ethics* (2005): 759–87; Arthur Ripstein, "Preference," in *Practical Rationality and Preference*, eds. Christopher W. Morris and Arthur Ripstein (Cambridge: Cambridge University Press, 2001), 37–55; H. L. Lillehammer, "Revisionary Dispositionalism and Practical Reason," *Journal of Ethics* 4 (2000): 173–90; Elijah Millgram, "Mill's Proof of the Principle of Utility," *Ethics* 110 (2000): 282–310, esp. 304–6; and, in explicit agreement with Enoch, Derek Parfit, *On What Matters*, vol. 1, 96 (Oxford: Oxford University Press, 2011).

concerns that does not especially reflect the agent's real evaluative perspective. For example, my own favored type of idealization will privilege favoring attitudes that are responsive to their object as it really is rather than as it is falsely imagined to be. Such attitudes quite plausibly are getting at what the agent really cares about. But that is not to say that all agents will agree that this method of idealization gets at their true concerns. Our subjectivist does not suppose that agents are infallible about the epistemology of their own genuine concerns. Thus they ought not turn over the question of the appropriate epistemic point of view where agent's genuine concerns are revealed to each individual agent. Rather, they must claim, the specified procedure is in fact well designed to get at the agent's genuine concerns even when the agent herself disputes this. And this opens up room, even within a subjectivist framework for the actual non-idealized agent to be alienated, in a sense, from her good.

Some have complained about such "full information" subjectivist views that they require stuffing a single head full of quite different sorts of lives and that there can be interaction effects which make it impossible to simultaneously appreciate what it would be like to live all the different sort of lives I might live.[6] Partly in response to such worries, other subjectivist views strive to avoid the need to stuff a single head full of what it would be like to live wildly different sorts of lives. They think it is enough to know the extent to which one's desires are satisfied in a life and to know the intensity of the satisfied and unsatisfied desires in a life, to know how well it has gone and to be able to compare how well it has gone to other lives one might have had.[7] Such views, like Benthamite Hedonism, maintain that the value of a life is "isolatable" and can be determined non-comparatively, whereas the more traditional full-information views suggest the normatively privileged attitudes are comparative between different ways one's life might go. This is why the traditional views have supposed we have to get all the options, as it were, before the agent simultaneously such that the agent might express preferences between them while fully appreciative of what each option is like.

But, and this is the important bit, both sides agree that the relevant favoring conative attitudes are accurately informed about their object. Subjectivists should focus, in one way or another, on desires of this sort.

[6] D. J. Velleman, "Brandt's Definition of 'Good'," *Philosophical Review* 97 (1988); D. Sobel, "Full Information Accounts of Well-Being," in my *From Valuing to Value*; C. Rosati, "Persons, Perspectives, and Full Information Accounts of Personal Good," *Ethics* 105 (1995); D. Loeb, "Full Information Theories of Individual Good," *Social Theory and Practice* 21 (1995).

[7] C. Heathwood, "The Problem of Defective Desires," *Australasian Journal of Philosophy* 83 (2005); E. Lin, "Why Subjectivists About Welfare Needn't Idealize," *Pacific Philosophical Quarterly* 100 (2019): 2–23.

Partners in crime?

Initially I said that subjectivists are those who accept the direction of explanation in the *Euthyphro* Dilemma that starts with attitudes. It might be more accurate to say that subjectivists are amongst those that champion the attitude first direction of explanation. Others, such as Kantians and Humean Constructivists, perhaps are best understood as doing so as well. Julia Markovits and Sharon Street, for example, share the subjectivist's main premises that value originates in our attitudes.[1]

Some Kantian views can be reasonably thought of as optimistic subjectivists who are persuaded that the procedurally ideal concerns of all possible agents will converge in a way such that all have decisive reason to be moral.[2] Such people think there is a non-moral mistake involved in immorality, such that even if we do not presuppose moral facts, procedurally excellent thinking will eliminate the desire to be immoral. Aside from issues I find with particular arguments to this conclusion, I find this view rather unlikely. Wouldn't it, for example, be quite surprising if necessarily there was incoherence or non-aesthetic mistakes involved in actions that produced aesthetically ugly things? Immorality seems mean and selfish, not necessarily incoherent and self-contradictory. Further, even if immoral actions are, surprisingly, necessarily incoherent, the normative importance of avoiding immorality feels quite distinct from the importance of avoiding incoherence and the explanation of what makes actions morally wrong seems quite distinct from the explanation of why they are incoherent.

[1] S. Street, "In Defense of Future Tuesday Indifference: Ideally Coherent Eccentrics and the Contingency of What Matters," *Philosophical Issues* 19(1) (October 2009); J. Markovits, *Moral Reason* (Oxford: Oxford University Press, 2014).

[2] See M. Smith, ibid. I offer reasons to doubt such convergence in my "Do the Desires of Rational Agents Converge?" in my *From Valuing to Value*.

Why subjectivism?

So much for what the view is, how to start putting it in its more plausible form, and who else might share such a view. It is time now to get to the central issue: Why should anyone believe it? I think the subjectivist should adopt at least a three-part strategy to motivate and defend subjectivism. Other components might be added. These possible additions will be discussed subsequently. But I think without at least these three, the view would not have a hope of being tolerably convincing. I will first outline the three phases and then offer some guidance about how I think the subjectivist should try to fill in each part.

In the first part, the subjectivist is on offense. Here she makes a case that there are contexts where subjectivism seems the most plausible theory. The view must show some areas or contexts where it has important advantages over rivals and that there are clear paradigmatic cases of reasons (or well-being) that are most plausibly determined by what we value. If there were never contexts where subjectivism had advantages over its rivals, why on earth would we accept it? So the subjectivist must point to cases where subjectivism looks like the most plausible approach. Ideally, they would make a convincing case that it is difficult to deny subjectivism this thin end of the wedge. Matters of mere taste, we will see, plausibly provide this sort of home ground for the subjectivist where their analysis is most tempting and difficult to resist.

In the next two parts, the subjectivist is on defense. In the second phase of her argument, she must confront non-moral cases where subjectivism seems intuitively implausible. Candidate cases of this sort are Parfit's example of someone who lacks any current motivation, even after procedurally ideal deliberation, to avoid future agony, Rawls's case of someone who wants to count blades of grass, or Gibbard's example of an ideally coherent anorexic who cannot stand being plump enough to sustain life. Sharron Street calls these characters "ideally coherent eccentrics."[1] These eccentrics are meant to highlight that there are cases where it seems we can rule out certain normative conclusions regardless of what the agent's motivational set looks like. If so, the direction that these eccentrics favoring attitudes point will not be normatively tempting. These cases put pressure on the ability of the subjectivist to extend her model beyond matters of mere taste.

Unlike in matters of mere taste, these cases suggest that there are attitude-independent constraints on what has value or is worth doing. If that were so, the range of cases where our attitudes determine what has value for us and what it makes

[1] Street, ibid.

sense to do would be severely constrained. The argument here against the subjectivist is in some ways similar to the argument in the *Euthyphro*. The thought is that there is a right answer in the cases being pointed to, and the correctness of that answer seems stable regardless of how we imagine the relevant attitude pointing. In this second phase, the subjectivist must convincingly respond to such cases and persuade us either that subjectivism can get the intuitive results or explain why the intuitions in such cases are less probative than might have been thought.

Finally, in the third phase, the subjectivist must confront the case of morality. What is morally required of an agent seems paradigmatically to not be determined by that agent's concerns yet to be normatively authoritative over her. The subjetivist about reasons cannot allow that both are the case. Thus, I think subjectivism is at least to this extent counter-intuitive and has some real explaining to do to earn our allegiance. Yet there are, I believe, a variety of considerations that can alleviate the pressure from morality on subjective accounts. (Note that the pressure being brought to bare here targets only subjectivism about reasons and does not seem to be nearly as serious of a problem for subjectivism about well-being.)

Other philosophers have defended subjectivism by adding a fourth phase: maintaining that subjectivism has epistemological and metaphysical advantages over its rivals or that our notion of a reason has features that could only fit with subjectivism. Typically the former thought is that subjectivism is better positioned to take advantage of the benefits of naturalism than its rivals. I won't pursue this strategy. Some of subjectivism's rivals have a solid claim to be compatible with naturalism.[2] So this strategy, even if it could vindicate these benefits of naturalism, would still need to vindicate subjectivism over its naturalistic rivals, and that, I maintain, would require something like the three-step program I outline. So this fourth step could not substitute for what I have outlined.

Bernard Williams famously pursued the latter thought above when he made a case that reasons must be capable of motivating, at least after proper deliberation. And since only desires can motivate, what one desires after proper deliberation must constrain what we have reason to do. But I think there are reasons that need not be capable of playing this sort of motivational role. Thus, I think this alleged short cut to subjectivism's neighborhood inadvisable.[3]

Other views, such as Parfit's and Scanlon's version of non-naturalism, are compatible with subjectivism about what grounds reasons even while being non-naturalist about what the reasons relation is. To be clear, Parfit and Scanlon both argue strongly

[2] For example, P. Foot, *Natural Goodness* (Oxford: Oxford University Press, 2001); R. Hursthouse, *On Virtue Ethics* (Oxford: Oxford University Press, 1999). For the response I think a subjectivist should give to such proposals, see David Copp and my "Morality and Virtue," in my *From Valuing to Value*.

[3] See Robert Johnson's, "Internal Reasons and the Conditional Fallacy," *Philosophical Quarterly* 49(194) (1999): 53–71 and my "Internalism, Explanation, and Reasons for Action in My *From Valuing to Value*. For a compelling contrary view, see Hille Paakkunainen, "Can There Be Government House Reasons for Action?" *Journal of Ethics & Social Philosophy* 12 (2017): 56–93.

against subjectivism. But as I read them, no part of that argument is their non-naturalism. Parfit is clear that even if you accept his sort of non-naturalism, subjectivism is still not ruled out. Subjectivism as I conceive of it is entirely compatible with non-naturalism at the metanormative level. I find subjectivism clearly more tempting as an account of what grounds reasons than as an account of what the reason relation is. Parfit and Scanlon rightly insist that until one has shown that what the reason relation is turns out to be nothing over and above some naturalistic state, one has not yet fully naturalized one's worldview. And thus even if we accept subjectivism about what grounds reasons, work remains to ensure a fully naturalistic view.

I have not found a route from considerations in favor of naturalism to a compelling case for subjectivism. But I also do not mean to assert that others who find such a case are mistaken. Let me plead division of labor here and not take a stand on the merits of such an additional component in helping to justify subjectivism. I will voice my suspicion, however, that such considerations have been generally overrated as justifications for subjectivism. However, the main point for now is that one need not reject non-naturalism to embrace subjectivism. Our central question is if, for example, one has a reason to eat chocolate ice cream because one has contingent conative favoring attitudes towards chocolate, and, if so, how generally it is the case that our contingent favoring attitudes ground reasons in this way. Our central question here does not concern whether naturalism is true.

The core thought behind subjectivism can remain whether we are naturalists or not. The core thought is that valuers generate value with their valuing. The direction of explanation goes from valuing to value rather than the reverse. That can be true whether we accept naturalism or not. Thus as I see it issues surrounding naturalism, epistemic worries, metaphysical worries, and whether subjectivists are identifying desire-satisfaction with what the reason relation is are possible additions to the subjectivist view, but not part of that core.

The thin end of the wedge for the subjectivist

Home turf for subjectivism, I submit, are matters of mere taste where we think there is not a compelling stance-independent one-size fits all answer for all.[1] If one happens to like flannel jammies more than cotton jammies, chocolate ice cream more than vanilla, or is more amused by David Chappell than Chris Rock, assuming one is fully and accurately aware of the non-evaluative features of these options, then I think it fairly clear one has a reason to go with the option one so favors over the option one does not and that this is made the case by the existence of this favoring attitude. Matters of mere taste seem, intuitively, to be cases where what matters to one really matters. Here, if anywhere, the road from valuing to value seems secure.

Peter Railton, in a famous passage, wrote:

> Is it true that all normative judgments must find an internal resonance in those to whom they are applied? While I do not find this thesis convincing as a claim about all species of normative assessment, it does seem to me to capture an important feature of the concept of intrinsic value to say that what is intrinsically valuable for a person must have a connection with what he would find in some degree compelling or attractive, at least if he were rational and aware. It would be an intolerably alienated conception of someone's good to imagine that it might fail in any such way to engage him.[2]

While I find Railton's words ultimately compelling, as a premise in a philosophical argument, I think one could reasonably complain that this was not common ground, that there were quite common intuitions that told against it, and that it was question-begging against a quite wide range of reasonable views.

But there is a scaled down version of Railton's claim that it is much less plausible to resist and that can more reasonably be treated as a compelling premise. There is, I maintain, a component of well-being (and normative reasons) that must find

[1] I make this case much more fully in my "The Case for Stance-Dependent Value," *The Journal of Ethics and Social Philosophy* (2019). See also Steven Wall and my "A Robust Hybrid Theory of Well-Being," forthcoming in *Philosophical Studies*.

[2] P. Railton, "Facts and Values," in his *Facts, Values, and Norms* (Cambridge: Cambridge University Press, 2003), 47.

this internal resonance with the person whose well-being or reasons it is. In matters of mere taste, such as choosing between patterns of dress or music or gustatory sensations, where intuitively we are choosing something because it is pleasing to ourselves rather than for other reasons, such resonance is critical to which such options benefit us. In such contexts, if it is to benefit me, options must resonate with me. I must in some sense favor or like it, at least if rational and aware. Call this the Minimal Resonance Constraint.

And, while I do think the Minimal Resonance Constraint is crazy intuitive, there are those that deny it. But this denial is not justified by bringing forward cases where intuitively our attitudes do not seem to ground reasons in the domain of matters of mere taste. It remains, I submit, highly intuitive that our attitudes ground our reasons in that domain. Instead, broadly speaking, the denial is motivated by finding cases outside of the realm of matters of mere taste in which the attitudes seem to lack authority and assuming that if attitudes lack authority in those contexts, they must also lack authority in matters of mere taste. But this crucial assumption, while warranted if the opponent were a full-blown subjectivist, is not warranted against the defender of the Minimal Resonance Constraint. Such cases outside the realm of matters of mere taste must eventually be confronted by the subjectivist, but they properly belong in the second and third stage, not this first one. Further, I maintain that when you kick the tires of the stance-independent attempts to capture our reasons in matters of mere taste, you notice that the problem they keep running into is a failure to heed our minimal resonance constraint (or to unconvincingly and without explanation try to mimic it). And I put it to you that you find that lack of resonance, at least in the context of matters of mere taste, unacceptable.

So, subjectivism has home turf. There is a range of paradigmatic and obvious reasons that it handles better than its rivals. That is a decent start on showing the view to be plausible, but it is only a start.

Agony, blades of grass, and all that

It seems possible to value things that are valueless and to fail to value valuable things. Some have offered cases that purport to show that this situation can remain even after the sort of procedurally excellent deliberation that subjectivists tend to accept. In the former category, we might remind ourselves of Parfit's example of a person who lacks any concern to avoid future agony or Rawls's example of a person who wants to count blades of grass. We think a person's future agony well worth avoiding, even if she does not now care about avoiding this agony after ideal procedural deliberation. We think counting blades of grass a waste of time and pointless. But the subjectivist says that whatever a person cares about in the right way makes what she so values valuable for her. In examples of this sort, the subjectivist is put on the defensive to explain why their view is not just defeated by such examples. In these cases, it feels like there are standards for what is valuable or disvaluable that is not hostage to what a person values even after procedurally ideal deliberation. Here I will briefly try to outline only how I think the subjectivist should respond to Parfit's version of this worry. The other examples would need independent treatment.

Parfit argues that we necessarily have reasons to avoid our own future agony. But seemingly one might fail to care now, even after ideal procedural deliberation, to avoid this future agony. Parfit admits that when one is experiencing the agony, one will necessarily mind it. But it does not follow, Parfit maintains, that a person must now care to avoid, even after ideal procedural deliberation, the future agony that one will later mind. Thus, the issue for Parfit involves the subjectivist being unable to vindicate the transfer of reasons we are confident exists between the reasons one will have later to get out of agony to reasons one has now to avoid it.

Why should we think this transfer of future reasons to current reasons is insecure on a subjectivist account? As Parfit points out, it is clearly psychologically possible that the knowledge that we will have a strong desire in the future, say to avoid hell, fails to produce a strong desire in the agent now to take steps to avoid that situation. We can as a psychological matter fail to be moved by the thought that something will matter to us in the future. Parfit is clearly right that descriptive psychology will not ensure the proper transfer of desires in cases where we are certain there is a transfer of reasons. In actual cases, the problem is usually that the future pain is, as Sidgwick put it, "foreseen but not fore felt".[1] The subjectivist suggestion that we provide agents with an accurate and retained impression of what the future agony will be like would

[1] Sidgwick, ibid.

surely go a long way to curing most actual cases of such irrationality. But Sidgwick thought that this would not solve all possible cases of such irrationality.

If Sidgwick is right, subjectivists cannot rely on the causal impact of accurate information about possible futures to ensure that agents are moved by their future concerns. I will grant this. Given that, the subjectivist can successfully respond to Parfit's challenge only by building in transfer principles into their account of ideal procedural deliberation that ensure that rational people will be moved today by the reasons they will have tomorrow. Parfit argues that, given their commitments, subjectivists cannot do this. He argues, "Subjectivists cannot claim that, if we were procedurally rational we would want to avoid all future agony."[2] I dispute this.

The subjectivist who claims that ideal procedural deliberation involves caring about one's future concerns is not assessing the content of one's future concerns and whether the objects of such concerns are worthy of being desired. Rather such a subjectivist principle is only concerned with whether one comes to care about the option as a result of an accurate understanding of it. The principle that one should now care about what one will later care about gives one no guidance until one starts to care about this rather than that for no good stance-independent reason. Such a view borrows no objectivist principles about what is worth caring about in the first instance. It is quite different from claiming that a person has a desire-independent reason to be moral or eat chocolate. The claim involves only the thought that if one will care about something later, one should now care about that fact. This seems continuous with the idea that one's passions set the ultimate goals and further reasons are hostage to what promotes our ultimate goals. Reason is still the scout or slave to the passions. Can it really be true that it is a distinctively anti-subjectivist principle that one should act so as to maximally comply with one's subjectively determined reasons over one's life?

Thus, I conclude that it is entirely open to the subjectivist to maintain that ideal procedural deliberation involves, among other things, caring about the future cares one will have after later ideal procedural deliberation. If this is so, the subjectivist can account for our necessarily having reasons to avoid future agony in an entirely natural way.[3]

Julia Markovits offers a different reply to Parfit's worry. She thinks that all agents, after ideal procedural deliberation, will have concerns that give them decisive reasons to be moral in all contexts. Markovits maintains that this apparently non-moral case about one's own future agony can be satisfactorily handled as a moral case. She maintains that since it would be immoral to pay no attention to one's own future agony, procedurally idealized agents will not be indifferent to such agony.[4] But, of course, this reply is only as good as the argument to the conclusion that all must care about morality after procedurally excellent deliberation.

---

[2] Parfit, ibid., 80.

[3] I argue for this conclusion more fully in my "Parfit's Case Against Subjectivism," in my *From Valuing to Value*.

[4] Markovits, ibid.

Even if Parfit's Agony Argument can be accommodated by the subjectivist, that would obviously only be the start of a successful defense at this second stage. What I hope to have gestured towards here is how the subjectivist can get started in replying at this stage. But I admit more work needs to be done here.

Much of the job will involve getting much clearer about the alleged counter-examples and fleshing them out more fully. We must keep in mind, for example, that in the grass-counting case, the agent fully understands exactly what all her options would be like and yet prefers counting blades of grass. This is quite different from a case where the agent wants to count blades of grass but when exposed to a life filled with friends and more meaningful activity, finds these latter sorts of activities more rewarding. Our agent never regrets her choice of counting blades of grass, even when vividly confronted with accurate understandings of what the alternatives would have been like. Those sorts of additions make me think it less clear that it really would be best for her to avoid counting blades of grass and return to the cocktail party. Rarely, I think, do people have in mind all the features that would be required to make the case a genuine problem for subjectivism. Nonetheless, I admit more needs to be said at this stage to fully defend the subjectivist view.

Reasons to be moral

To my mind, the most compelling case against subjectivism flows from the thought that morality provides everyone with some significant reason to obey even if a person lacks any contingent concerns that would be furthered if they obey. That is just to repeat the often-made observation that morality is categorical rather than hypothetical and one cannot escape the force of the moral requirement simply by failing to care about it.

It seems that everyone has a significant reason to not, for example, abuse the vulnerable – say steal the gloves and shoes of a homeless person on a very cold night. But subjectivism cannot ensure such a verdict, and so, the argument under consideration here goes, we should reject subjectivism. Call this the Amoralism Objection against subjectivism.

Some subjectivists resist the claim that they are in the pickle I describe. Mark Schroeder holds out real hope that even within a subjectivist framework, all agents will necessarily have most reason to be moral.[1] Thus, his subjectivist view, he thinks, might generate results similar to those the Kantian expects. In Schroeder's case, this hope is tied to maintaining that the weight of reasons is not directly responsive to the strength of desire or the degree to which an action promotes something we desire.[2]

I think the subjectivist does have to deal with possible worlds where procedurally idealized agents lack decisive reason to be moral. But I think the subjectivist has a surprisingly compelling explanation for our intuitions about such cases that is compatible with subjectivism. I cannot outline the whole story here, but I can provide a taste. Consider first that a great many people, most polls suggest over 95 percent in the United States, think there is at least a decent chance that God exists and that the quality of one's afterlife is positively affected by the moral acceptability of one's life. If this were true, one would have extremely powerful subjective reasons to live a moral life. As Pascal's Wager showed us, all it takes is the belief that being good on earth has at least a tiny probability of making our afterlife better than it would have been than if we had been bad to make it rational to invest substantially in avoiding that extra chance of a bad eternity. If most people think being good on earth improves one's prospects after one's death, yet this claim is false, we would have a strong debunking explanation for why so many find it intuitive that one has a strong reason to be moral regardless of one's concerns.

[1] M. Schroeder, *Slaves of the Passions* (Oxford: Oxford University Press, 2007).

[2] I take issue with Schroeder's arguments to this position in my "Subjectivism and Proportionalism",

Second, as demonstrated by Plato's example of the Ring of Gyges, there are many robust reasons to be moral for those of us that lack the ring that have to do with serving our concerns. Detected immorality has a strong and not very coincidental tendency to harm one's prospects whether by resulting in being incarcerated or just shunned from the benefits of mutual cooperation and friendship.

Third, most people would significantly prefer to get the goods that tempt them towards immorality without harming others. Being forced to harm innocent strangers, for example, is, by the lights of the vast majority of people's concerns, a cost. Most people resonate in common with humanity and do not intrinsically want to harm many others. This would ground some reason to avoid immorality. One might claim that it would not ground a decisive reason to be moral. That is true, but it is not at all clear to me that common sense supports the existence of such in all cases. I think common sense suggests that all have a significant reason to obey morality but does not insist that all always have decisive reasons to obey. The considerations put on the table would go a very long way to explaining in a subjectivist friendly way the existence of such an intuition.

Fourth, the subjectivist can allow that all necessary ought to be moral, when the 'ought' is given a moral reading. What the subjectivist denies is that necessarily all ought to be moral when the 'ought' is given a rational or reasons reading. This ambiguity can help explain why it seems intuitive to say that all necessarily ought to be moral.

Fifth, there is a powerful lore that evil does not pay. That is, in the words of our amazing local graffiti artist, the bullets you send will meet you in the end. This lore has many legitimate sources, such as some of what I mention previously, and seems true enough to make the parental policy of teaching one's children to be nice and share their toys make good sense. But this lore is propagated via stories beyond what I think most would agree is true. We rarely tolerate in our more popular fiction evildoers who do not get their comeuppance in the end. This is especially true of stories designed for younger, more impressionable minds. We insist that children be taught that evil does not pay and highlight the (literal) ugliness and folly of such behavior. This encourages the conflation of the thought that there is enough truth in such stories to matter and reasonably to guide parents in how to raise their children with the unlikely idea that evil cannot pay.

I could go on in this vein.[3] But enough has been said to hint at how the subjectivist can perhaps convincingly respond to the Amoralism Objection. Broadly, we should be unsurprised if we have to give up some aspects of common sense in coming up with the most plausible general theory of reasons for action. The intuitive costs of accepting subjectivism stemming from the Amoralism Objection, I have offered some reason to think, are not as dramatic as they initially seemed. It is surprisingly plausible that

in my *From Valuing to Value*.

[3] And indeed I do in my "Subjectivism and Reasons to be Moral," in *From Valuing to Value*.

we can explain the appearance of categorical reasons to be decent by appealing to a combination of very, very robust hypothetical reasons to do so and mistaken beliefs. And there are significant considerations in favor of subjectivism that plausibly can go some way towards outweighing what intuitive costs remain from the Amoralism Objection.

Conclusion

In conclusion, in my view, most of the case for subjectivism has to hinge on showing three things. First the subjectivist must persuade us that there is a range of cases that subjectivism handles more convincingly than its rivals. I think the realm of matters of mere taste is the most fertile ground for the subjectivist here in both the case of well-being and reasons. Next, any defense of subjectivism should acknowledge that there are a range of cases in which even procedurally idealized conative attitudes seem like they can hit on the wrong answer. This has been alleged to happen both inside and outside the realm of morality. And the subjectivist's attempt to blunt the force of such alleged counterexamples in these two realms constitute what I think of as the second and third arena in which the subjectivist must successfully duke it out with its rivals to adequately defend their view. Often the starting point for the subjectivist's response in these latter two arenas will be to suggest that as a result of almost all actual people robustly having various concerns, we get fooled into ruling out anyone having intrinsic reasons to, for example, eat feces. We tend to not waste our time thinking very hard or seriously about what life would be like for a person with wildly unfamiliar sorts of concerns or what such a person would have reason to do. I am optimistic that moves of this sort will prove ultimately compelling in blunting the force of the cases alleged to be counterexamples to subjectivism both inside and outside the realm of morality. Finally, it must be remembered that subjectivism's rivals often bump up awkwardly against our intuitions as well and have other difficulties. We ought not fixate on the blemishes on subjectivism and hastily reject the view for we may find that our other options are even more bruised.

22. Kantian constructivism

1 Introduction

In most corners of philosophy, one can find a view called "constructivism". Constructivists about mathematics think that proving the existence of a mathematical object means constructing it according to some specified procedure. Social constructivism is the view that the nature of a thing – mental illness, say – depends in significant ways on its position within a social structure. Constructivism about normativity is the view that normative facts – some of them, anyway – are built out of the attitudes and activities of a specified set of agents by a procedure reflecting the dynamics of practical reason.

These programs are united by two ideas. The first is a qualified anti-realism about their target. To say that something is constructed is to say that it does not exist independently of the basis out of which it is constructed. This isn't to say that it doesn't exist at all, or even to deny it "full-blooded" existence, but merely to deny its existence is *sui generis*. The second is that this dependence is somehow subtle, sophisticated, or otherwise non-obvious. This is how constructivism avoids the pitfalls of more simple-minded anti-realisms. Mathematical constructivism is not the view that the only mathematical objects that exist are those identified by flesh-and-blood mathematicians. It is the view that the only objects that exist are ones that can be constructed according to an idealization of those mathematicians' procedures. Similarly, no constructivist about normativity would hold that June's liking a smelly shoe suffices, all by itself, for the shoe to be valuable, or that it gives June a reason to keep the shoe. June's preference must be ratified by some process of rational scrutiny. This sophistication gives the constructivist the resources to better capture our considered judgments about mathematics, social reality, and normativity than a more artless anti-realism. It enables them to explain how the discovery of certain facts – the infinitude of prime numbers or the evil of torture – represent genuine achievements rather than navel-gazing.

2 Motivations for normative constructivism

Constructivism about normativity has some unique attractions beyond the generic ones just mentioned. First, it seems to avoid some of the metaphysical problems that normative realism faces. If we are satisfied with the metaphysics of our construction basis, and we maintain that the normative facts in our target class are nothing "over and above" this basis, then we ought to be satisfied with the metaphysics of these facts as well. Constructivists differ in their ambition on this score. Some aspire to show how normative facts *tout court* can be constructed out of non-normative materials.[1] Others are content to take one domain of normativity for granted in giving a constructivist gloss on another, usually more mysterious, domain. They might, for instance, uncritically employ facts about rationality in constructing moral or political facts.[2] Second, the dependence of the normative on our attitudes and activities that the constructivist proposes eases the threat of alienation. If normative facts are in some sense a reflection of the activity of our own practical reasoning, operating on our own antecedent concerns, then we are less likely to be "left cold" by them than if they were grounded in lifeless parts of the world. Third, constructivism seems to capture the practicality of the normative. In construing normative facts as the result of a certain activity, it recognizes that, as Korsgaard puts it,

> normative concepts are not ... the names of objects or of facts or of the components of facts that we encounter in the world, [but] the names of the solutions of problems, problems to which we give names to mark them out as objects for practical thought.[3]

It's worth saying a bit more about these last two advantages, since some critics think they are not advantages at all. Constructivists have been accused of confusing motivating and normative reasons.[4] Sure, some critics say, a normative principle that

[1] Sharon Street, "Constructivism about reasons", in Russ Shafer-Landau (ed), *Oxford Studies in Metaethics*, vol. 3 (New York: Oxford University Press, 2008), pp. 207–45.

[2] Julia Markovits, *Moral Reason* (New York: Oxford University Press, 2014); John Rawls, *Political Liberalism* (New York: Columbia University Press, 2005), pp. 89–129.

[3] "Realism and constructivism in twentieth century moral philosophy", reprinted in *The Constitution of Agency* (New York: Oxford University Press, 2008), p. 322.

[4] See, for example, Derek Parfit, *On What Matters*, vol. I (New York: Oxford University Press,

did not reflect an agent's own evaluative attitudes would provide that agent with no *motivating* reason to comply with it, and so might not bind her, psychologically speaking. But that doesn't mean that such a principle would not bind her in the sense of presenting considerations that *ought* to move her. Perhaps coldhearted Craig will be unmoved to help suffering Sarah. Nonetheless, Sarah's suffering is a normative reason (a moral reason) for Craig to help.

This criticism is misguided. The constructivist's aim is to explain the significance of the long and imperious word used so freely in this objection – to explain what makes something "normative". To call something normative seems to suggest that it has a special "hold" over us. Why is it that certain sorts of considerations have this "hold", while others – the edicts of a bogus religion or the charges of the fashion police – do not?

When someone tells us that we've gone wrong morally or that we're behaving self-destructively, these charges seem to have a claim on our attention that "you've committed a fashion crime" does not. Fashion rules certainly *apply* to us: they are about us, and may be addressed to us.[5] But this is, as Philippa Foot has put it, merely a "piece of linguistic usage": we are part of the domain to which they purport application.[6] From this linguistic fact nothing about the normative authority – the *hold* – of these rules follows.

For the realist, the "hold" that normative facts have over us will be something of a brute fact. It just is the case that, for example, the moral law has it and *Dianetics* doesn't. Someone who disputes the authority of these normative facts will find the realist's insistence on that hold to be little more than the begging of questions and the stomping of feet. The constructivist, by contrast, offers an explanation of what this "hold" comes to – an explanation that is designed to persuade the very person whom those normative facts purport to bind. Normative facts reflect the evaluative perspective of that agent and are constructed through the activity of practical reason out of the attitudes and commitments of that agent. This connection means that these facts will "stick" to the agent in more than just a linguistic sense because they *matter* to the agent. Whether this "hold" will succeed in moving the agent remains, of course, an open question. Constructivist reasons, after all, reflect not the actual motivations of the agent but rather those that can be constructed out of her actual concerns through the activity of practical reason. The agent may well, due to shortcomings of rationality, fail to be motivated accordingly.

2014), p. 66. For more on this charge in Parfit, see Julia Markovits, "On what it is to matter", in Simon Kirchin (ed), *Reading Parfit* (New York: Routledge, 2017).

[5] Cf. Bernard Williams, *Moral Luck* (Cambridge: Cambridge University Press, 1981), p. 122.

[6] "Morality as a system of hypothetical imperatives", reprinted in *Virtues and Vices* (New York: Oxford University Press, 2002), p. 160.

3 Varieties of normative constructivism

Some constructivists insist that these advantages – metaphysical conservatism, non-alienation, practicality – come with costs. For constructivism to offer these benefits, it must take the agent's actual evaluative attitudes as its construction basis and a thin, procedural conception of practical reason as the construction procedure. But given the wide variation in people's evaluative attitudes, such a view threatens the universality and objectivity of moral and prudential facts.

This suggests a relativism that can take one of two forms. We could say that the content of a particular normative domain, like morality, is relativized to particular agents and their evaluative attitudes. Thus morality-for-Jeff and morality-for-June could be composed of distinct normative facts depending on Jeff's and June's evaluative attitudes. If, for example, Jeff simply did not care about the interests of others, even when fully coherent, there would be no obligation to help others in morality-for-Jeff. On the other hand, we might say there is just one morality, but its authority is limited. Here Jeff might be morally obligated to help others, but this would have no more "hold" over him – entail no more about his reasons – than the canons of Scientology. It is worth giving up on such claims to the universal authority of morality, these constructivists think, for the sake of rendering moral facts metaphysically respectable, non-alienating, and practical.

Constructivists who take either of these roads are often called "Humeans". This is a nod to two Humean doctrines: a strong connection between moral judgment and contingent "sentiments" and skepticism about the powers of practical reason. Hume endorses a formal (or procedural) conception of practical reason, according to which reason tells us how to deliberate about, or with, our ends, desires, or evaluative attitudes but doesn't tell us which of these to adopt in the first place.[1]

Bernard Williams is a constructivist of this Humean sort.[2] He writes that for some agents who are simply unmoved by moral considerations, even after we reason with them, because they lack the relevant evaluative attitudes, there may be no further resources available by means of which to somehow "stick" our moral "ought" claim to the agent; nothing, that is, "except the rage, frustration, sorrow, and fear of someone who sees someone else convincedly or blandly doing what the first person morally

[1] *A Treatise of Human Nature*, book II, part III, section III, paragraph 6.

[2] So is "Constructivism about reasons".

thinks they ought not to be doing." In this instance, Williams says, "this critic deeply wants this *ought* to stick to the agent; but the only glue there is for this purpose is social and psychological."[3] This glue, he suggests, is all we should be looking for. The issue is not whether our wrong-doer has normative reasons to act better but whether we can somehow trigger his reformation.

Kantian constructivism can be understood as a response to this view. Kant (and Kantians) share Hume's (and Humeans') preference for a formal conception of practical reason. They are skeptical about the "dogmatic" use of reason, in particular the claim that moral facts can be apprehended by reason in the same way that geometrical facts (allegedly) can be. And like Humeans, Kantians are moved by the attractions of metaphysical tractability, non-alienation, and practicality. What distinguishes Kantians is their refusal to concede that these views force us to give up on a universally authoritative morality. On the contrary, many Kantians are attracted to constructivism precisely because they think it is our only hope of establishing the universal authority of morality.

Kant seems to understand universal authority to be part and parcel of our concept of morality. He begins his inquiry by asking what the "supreme principle" of morality so understood could be. Previous efforts at this endeavor, he argues, were bound to fail. Either (like Hume) they made the content of the moral law dependent on the idiosyncratic desires of the individual agent, in which case all hope for universality was lost. Or else (like dogmatic rationalists) they declared the law to be entirely independent of the wills it claims to bind. This, Kant thought, amounted to much the same thing, since a moral law that did not arise from the will of the agent could bind that agent only if the agent had some desire or incentive to comply. This brings us back, in other words, to the two forms of moral relativism described previously: a desire-relativism that is built into the content of morality itself or a morality that may be *about* all rational agents but which has no more normative authority over them than the prescriptions of Scientology.

Far from seeing relativism as the inevitable upshot of a constructivist approach to ethics, Kantian constructivists see the answer to the moral relativist as *requiring* a constructivist approach. Kant saw an alternative to the Humean and "dogmatic rationalist" approaches. The supreme principle of morality, he argued, is somehow inherent in the idea of a rational will *as such* and will therefore be universally authoritative not because of some contingent convergence of evaluative attitudes, nor because of the authority of some fact external to the will, but because it is built into what it is to have a practical point of view. In other words, only by recognizing that "the human being ... is bound only to act in conformity with his own will, which however ... is a will giving universal law"[4] can we avoid both horns of the relativist's dilemma.

---

[3] B. Williams, "*Ought* and moral obligation", in *Moral Luck* (Oxford: Oxford University Press), p. 122.

[4] *Groundwork for the Metaphysics of Morals*, 4:432–3.

All this suggests a constructivist project. Discovering the "supreme principle" of morality means discovering a principle that can be constructed by practical reason *as such* from whatever heterogeneous material humanity may furnish it with, and thus with a principle that will "stick" to agents no matter their evaluative starting points. This project, Kantians believe, is the key to a defense of moral principles that are universally binding by agents' own lights. If they are right, then the possibility of a moral law that is universal in form but not *alienated* from the agents it claims to govern will depend on a demonstration. It will depend on our showing how a merely formal conception of practical reason can guarantee convergence on a single principle by all rational creatures as they shift through their endlessly diverse sets of evaluative attitudes.

This is an ambitious project, and, as always, the proof of the pudding is in the eating. Unsurprisingly, the doctrines Kantian constructivists are most interested in constructing resemble parts of Kant's own ethical system. In the following sections, we canvass arguments dealing with different formulations of the moral law.

4 Constructing universalizability requirements

Kant's first formulation says that you should only act on practical principles that you can also will to be universal laws. There is a relatively straightforward constructivist gloss on Kant's argument for the universal normative authority of this Formula of Universal Law.[1] Every practical reasoner faces the problem that is distinctive of practical reason, what amounts to a practical version of the problem of free will. Whether or not we are free in any deep, metaphysical sense, we must act under the assumption that we are free. But action is a causal process: my decision to wiggle my toe *causes* my toe to move, and that has knock-on effects throughout the whole causal order of the universe. My action must be governed by laws because all causal processes are. But these two requirements seem to push in opposite directions, or, at the very least, they pose a problem that Kant thinks is characteristic of practical reason. How can I act so that my action is both law-governed and free?

Suppose I make my "supreme principle of practical reason" – the principle that I employ in evaluating the suitability of all other principles – something that directs me to respond in a specified way to a specified object or end – to pleasure, a piece of zucchini, or Schubert's Trout Quintet. If my behavior is law-governed, then this principle would have to also subsume the behavior of all free and rational creatures. But this dubious. Not every rational creature will be naturally inclined to respond to zucchini or the Trout Quintet in the way I do. It's unlikely to be a fully general law about the activity of rational creatures, and it certainly isn't a law that reflects the *free* action of all rational creatures. So the idea of a principle organized around a particular object is a non-starter.

What does that leave us with? Well, Kant explains, the only suitable principle will be one that guides our action through its "law-giving form", that is, *not* one that mentions a particular object but one that requires no more and no less of us than to act on principles that *could be laws*. And that's just what the Formula of Universal Law requires; it says that our principle of action must be "universalizable". Thus this imperative has normative force for all rational creatures because it is the only adequate

[1] The clearest version of this argument appears in the *Critique of Practical Reason*, 5:19–5:35. The reconstruction offered here takes significant interpretative liberties. On Kant's own views vis-à-vis constructivism see Patrick Kain, "Realism and anti-realism in Kant's second *critique*", *Philosophy Compass* 1(5), 2006, pp. 449–65.

solution to the characteristic problem of practical reason, the problem of marrying freedom and lawfulness.[2]

What does this show? Even setting aside cavils about the argument itself, we can wonder how much distinctively moral content falls out of it. Hegel famously calls the Formula of Universal Law an "empty formalism", and it's easy to see why. Since the argument only incorporates other people in a formal way – as things falling within the scope of a universal quantifier – it is mysterious which forms of interpersonal conduct the law will enjoin and forbid.[3] Kant does claim to derive particular duties from the principle by arguing that the routine violation of such duties (as if by law) would be either impossible or self-defeating. But his arguments are contentious, and his application of the principle invites a swarm of dubious results. Perfidy seems permissible so long as it is undertaken with hyper-precise principles. And benign plans may turn out to be condemned because of mundane coordination problems.[4]

For these reasons, many Kantians think that the Formula of Universal Law requires supplementation or amplification by other parts of Kant's system. Let's turn to those.

[2] Cf. Christine M. Korsgaard, *Sources of Normativity* (New York: Cambridge University Press, 1996), pp. 97–8.

[3] On this objection, see Sally Sedgwick, "Hegel on the empty formalism of Kant's categorical imperative", in Stephen Houlgate and Michael Baur (eds), *A Companion to Hegel* (Malden: Wiley-Blackwell, 2006), pp. 265–80.

[4] For a summary of these problems from the point of view of someone skeptical of the centrality of the Formula of Universal Law to Kant's ethics (and of the idea of "Kantian constructivism"), see Allen W. Wood, *Kantian Ethics* (New York: Cambridge University Press, 2008), pp. 71–3.

5 Constructing respect for persons

The Formula of Humanity may be a more promising focus for constructivists. Kant says that all persons possess an "unconditioned, incomparable worth". This "dignity", as Kant calls it, merits a distinctive form of regard: *respect*. Most fundamentally, this means we must recognize the authority of other persons to make valid claims on us. More specifically, it forbids us from treating others as "mere means" – as things we can employ for our own purposes in complete indifference to their own capacity for rational choice. And it requires us to regard all persons as "ends in themselves". (An "end" for Kant is anything for the sake of which we act, and thus a broader category than the more familiar notion of the goal an action aims to achieve. Humanity can be an end in this broader sense insofar as we can act for its sake by paying it the appropriate respect.)

Kant says that the Formula of Humanity is a requirement of practical reason as such. His argument proceeds by investigating the way that some values are conditioned on others.[1] It begins with optimism that some of the ends we pursue are ones we have *reason* to pursue:

> 1 I value the ends I rationally set myself and take myself to have reason to pursue them.

It then appeals to a constructivist-flavored premise:

> 2 But I recognize that their value is only conditional: if I did not set them as my ends, I would have no reason to pursue them.

But, Kant asks, why think that we can generate reasons to promote some end just by adopting it? We must, he says, think that we have the power to confer value on our ends by rationally choosing them:

> 3 So I must see *myself* as the condition on the value of my ends – as having a worth-bestowing status.

[1] The discussion in this section closely follows the discussion in J. Markovits, *Moral Reason* (Oxford: Oxford University Press, 2014), especially §5.4. Compare the (highly compressed) remarks at *Groundwork*, 4:428–9. For more on this argument as a reading of Kant, see C. M. Korsgaard, "The formula of humanity" reprinted in *Creating the Kingdom of Ends* (New York: Cambridge University Press, 1996), pp. 106–33; Allen W. Wood, *Kant's Ethical Theory* (New York: Cambridge University Press, 1999), pp. 124–

From this, Kant seems to infer that we must accord ourselves *un*conditional worth:

> 4 So I must see myself as having an unconditional value – as being an end in myself and the condition of the value of my chosen ends – in virtue of my capacity to bestow worth on my ends by rationally choosing them.

But at this step, I should also recognize that the same argument holds from your perspective, concerning *your* rational nature, and so consistency requires that I attribute the same worth-bestowing status, and so the same unconditional value, to *you*, and to *any other rational being*:

> 5 I must similarly accord any other rational being the same unconditional value I accord myself.

Hence the Formula of Humanity:

> 6 I must always act in a manner that respects this unconditional value. I must treat humanity, whether in my own person or in the person of another, always at the same time as an end, never merely as means.

As it stands, this argument raises significant worries. Rae Langton, for example, confesses a temptation to describe it as "a chain of non sequiturs."[2] What's irrational – more specifically, irrational in the *thin, procedural* sense the constructivist relies on – about simply taking each of my ends to be valuable in itself, unconditionally, and independently of my having chosen to pursue it? And even if I concede that my ends' value is somehow *conditional on me*, why conclude from this that I must have *unconditional*, intrinsic value? Not all sources of value are themselves valuable, much less intrinsically so. Infection makes penicillin valuable, but isn't itself valuable; the cubic press, which turns graphite into diamonds, makes carbon valuable, but is itself only instrumentally, not intrinsically, valuable. Are we really to conclude that we're valuable only in the way that the press is valuable – because we turn lumps of valueless world, like lumps of graphite – into the good stuff? And it is far from clear, given what Kant has said, that *I* (rather than something else) must be the *ultimate* source of value of my ends, even if we concede that the source of their value *is* intrinsically valuable. And even if I *am* the intrinsically valuable source of value of my ends, what commits me to thinking *you* are an intrinsically valuable source of value, too?

Let's look closer at the value dependency claim driving this argument. First consider an ordinary instrumental imperative. If you want good dental health, floss regularly. It would be irrational to value the end of good dental health, but not value regular flossing. The value of the more fundamental end implies the value of the instrumental

32; Jens Timmermann, "Value without regress", *European Journal of Philosophy* 14(1), 2006, pp. 69–93.

[2] Rae Langton, "Objective and unconditioned value", *Philosophical Review* 116(2), 2007, p. 169.

end. Kant's argument suggests that the reverse implication may also hold: the value of an instrumentally valuable end implies the value of the more fundamental end to which it is instrumental. It would be irrational to value regular flossing without valuing good dental health (in the absence of other reasons for flossing). It would be equally irrational, Kant's argument suggests, to value my contingent (non-instrumental) ends without valuing the source of *their* value – the value of the rational nature that set them. If I'm rational, I'll value flossing because I value good dental health because I value pain prevention because I value *me*.

But why think that I am the only possible source of value of my contingent ends? Why can't I, rationally, just take them to be valuable in themselves, unconditionally? Let's start with the easier case: imagine a person who, when asked why he flosses regularly, responds that he does it for its own sake. And imagine that he gives a similar response when we ask him why he does all the other things he does. Such a person's value commitments would strike us as bizarre, in large part because of their total lack of internal coherence. There's just something arbitrary and dogmatic about valuing so many unrelated, unsystematic, contingently chosen ends, without some more fundamental explanation for why they matter. Compare the epistemic case: a person who, when asked why she believes each of the things she believes, responds, "I just *do*." Rational people's sets of beliefs are not so piecemeal and disconnected. Their beliefs cohere and support each other. Justification may have to bottom out somewhere, but it had better not bottom out in *too* many unrelated articles of faith – especially not articles of faith about which there is intractable disagreement between otherwise rational agents.

One advantage of valuing humanity as an end in itself, and recognizing it as the source of the value of my other ends, is that it can bring *systematic unity* to my ends. A set of contingent ends that includes the end of humanity is rationally preferable to one that does not because it is, to borrow a term from Michael Smith, more "systematically justified." For Kant, the ideal of systematic unity – of having our ends or beliefs stand in a network of mutually supportive, reciprocally justifying relations – is one the principal aims of reason in both its practical and theoretical employments.[3] Systematic unity is a very demanding (perhaps unreachable) ideal, but it is nonetheless rooted in a procedural conception of rationality. It's a matter of my ends' (*inter alia*) standing in the right *relations to each other*, rather than of my having or lacking a particular end. (In a sense, this is the crux of Kant's reply to Hume and the dogmatists: a procedural conception of reason can still be very demanding, if its demands are sufficiently schematic.) If this is right, then there is rational pressure on us, as Kant thought, to search for "an unconditioned condition" of value – an answer to the string of why-questions we might ask about the value of the things we care about.

[3] For a treatment of the idea that delves into Kant more deeply than we can here, see Karl Schafer on "Rationalist Kantian Constructivism" in "Realism and constructivism in Kantian metaethics", *Philosophy Compass* 10(1), 2015, p. 706.

But the argument so far cannot explain on its own why it's procedurally irrational to trace the chain of value-dependency among our ends back to a *different* starting-point. Many ends, it seems, would increase the coherence and systematic justifiability of our set of ends if we came to see them as the source of value of those ends. As Parfit observes:

> Consider ... Smith's claim that we can be rationally required to have a more unified set of desires. Mere unity is not a merit. Our desires would be more unified if we were monomaniacs, who cared about only one thing. But if you cared about truth, beauty, and the future of mankind, and I cared only about my stamp collection, your less unified set of desires would not be, as Smith's claim seems to imply, less rational than mine.[4]

Parfit's point illustrates that not any kind of unity of ends is, intuitively, equally rational. Reasons judgments lay claim to a validity that is *non-parochial* – that can be recognized from *any* perspective. If we begin, as Kant says we do, from an optimism that that some of the things that matter to us *really matter* – that we have genuine reason to pursue and protect and respect and promote them – then we are claiming more for our ends than just that they're *what we're after*. In this way, our ends resemble our beliefs: if we take our beliefs to be rational, then we take them to be justifiable in a way that *others*, or we ourselves, should our preferences later change, should be able to recognize; we're not merely saying they're what we happen, now, to think.[5] This goal of non-parochialism turns the search for systematic *unity* among our ends into a search for systematic *justifiability*.

The possibility of intrapersonal changes of heart about value pushes us in the same direction. One of the goals of the constructivist project is to identify a conception of value that is non-alienating. The stamp collector may not be alienated now from a conception of value that identifies stamps as the ultimate source of value. But should her values change, she would certainly find herself alienated. Much better, then, to trace the value of her ends back one step further: to her own will. Her own will, after all, is something from which she cannot become alienated.

So much, then, for stamp-collecting. It doesn't even provide stable systematic justification to *our own* ends, much less make sense from the perspective of anyone else's. One of the main advantages of the constructivist conception of reasons, we suggested earlier, was that it seems less *dogmatic* than realist views that posit "external" normative facts that are completely independent of the agent's perspective. But insisting that stamp-collecting is an ultimate worth-bestower is *very* dogmatic. It totally dismisses most other people's perceptions of value from the start, with no way of defending the dismissal. So it's important that the end we recognize as the source of value of our

[4] *On What Matters*, vol. I, p. 80.

[5] Smith embraces the same standard in "Internal reasons", *Philosophy and Phenomenological Research* 55(1), 1995, p. 118.

own ends – and the linchpin of their systematic justification – makes sense as *potential* source of value for the ends of others, or, indeed, our own ends later on should our values change.

Stamp-collecting is, of course, not the only, or most plausible, alternative systematizer of value. *Happiness* seems like a good (and philosophically popular!) candidate. Perhaps we should think our ends are valuable not because we choose them, and we're unconditionally valuable, but because they make us happy, and happiness is unconditionally valuable.

Taking happiness to be the "unconditioned condition" of value makes pretty good sense of most of my commitments and of many of the commitments of others. But despite the importance almost everyone attaches to happiness, it cannot, it seems, explain the value we attribute to *all* our ends. Many people value ends quite independently of whether they generate happiness. So the assumption that happiness is "the source of value" will still force us to dismiss many value-commitments out of hand. The transcendental hypothesis that persons' capacity for valuing is both source and conferrer of value fares better: it allows us to begin with the default assumption that everyone's ends matter and correct that assumption only when it actively conflicts with the commitment to the value of humanity.

The goal isn't, of course, to find an ultimate end that will accommodate *everything* individual people happen to value. The point of a moral principle, after all, is partly to correct our value judgments. But it shouldn't dogmatically rule out some people's values as mistaken from the start. We should grant anyone's ends, not just our own, the benefit of the doubt, as a kind of working assumption, and correct that assumption only when we need to. This at least is the goal and appeal of the constructivist project, as we have interpreted it. If we assume that people are the source of value, then their value can, at a first pass, explain the value of *any* chosen end, though that end could later turn out to be irrationally adopted if it (or its pursuit) necessarily conflicted with respect for the special value of persons.

This argument explains why there is rational pressure on all of us to value humanity as an end, regardless of our contingent ends and commitments. And so it provides the first necessary component of a successful constructivist defense of the thesis that rationality requires us to be moral. But the argument also provides a second necessary component of such a defense. It explains why the rationally required end of humanity is *not just one end among others* but *trumps* those others in cases of conflict and so can be a source of moral *requirements*. Because the value of humanity is, in the view we've sketched here, a condition of the value of any other end whatsoever, it is always procedurally irrational to fail to treat it as an end for the sake of promoting some particular end-to-be-effected. Thus Kant's moral imperative can never be overridden by instrumental or prudential concerns. Even in a constructivist view of practical reasons, we always have most reason to do as morality requires.

6 Constitutivist supplements

Constructivism is frequently supplemented by claims about the "constitutive" nature of the entities involved in normative construction. There are two reasons this supplementation may be necessary. The first is a problem for more ambitious constructivists who want to give a constructivist analysis of normative facts quite generally (i.e. for Street but not Markovits or Rawls). The procedure that a constructivist suggests yields normative facts is itself normative: there is a right way and a wrong way to go about constructing facts about reasons (for example) from facts about evaluative attitudes. But what is the status of these normative facts? What is it, in other words, that makes one procedure of construction appropriate rather than another?[1] The second problem is more basic. It is the simple fact that, arguments from the previous section notwithstanding, Kantian constructivism is a hard road to hoe. Because practical reasoning as such appears to be such a mutable activity, it is hard to see how we could guarantee that much of anything will be inevitably endorsed by all practical reasoners, much less a demanding moral doctrine like the Formula of Humanity.

These two problems have led a number of constructivists to supplement their arguments in similar fashion.[2] The idea is that the states involved in practical reasoning have a constitutive nature and that from the point of view of practical reasoners, this constitutive nature has a special normative authority. If this is right, then these constitutive facts may be a source of normative constraints on our construction procedure. Street, for example, identifies constitutive features of taking oneself to have a reason:

> Just as it is constitutive of being a parent that one have a child, so it is constitutive of taking oneself to have conclusive reason to Y that one also, when attending to the matter in full awareness, take oneself to have reason to take what one recognizes to be the necessary means to Y.[3]

This constitutive feature, she goes on to say, has normative implications. If you take yourself to have conclusive reason to Y, it is correct for you to also take yourself to have

[1] For objections turning on this question, see Russ Shafer-Landau, *Moral Realism: A Defense* (New York: Oxford University Press, 2003), pp. 44–50; David Enoch, "Can there be a global, interesting, coherent constructivism about practical reason?" *Philosophical Explorations* 12(3), 2009, pp. 313–39; Nadeem Hussain, "A problem for ambitious metanormative constructivism", in J. Lenman and Y. Shemmer (eds), *Constructivism in Practical Philosophy* (New York: Oxford University Press, 2012).

[2] For more on why these views are "natural bedfellows", see "Realism and constructivism in Kantian metaethics", pp. 691–2.

[3] "Constructivism about reasons", p. 228.

a reason to take what you recognize is the necessary means to Y. The construction procedure that Street advances is grounded in claims about what is constitutive of holding an evaluative attitude.

It seems unlikely that the constitutive requirements of *valuing* as such will get Kantian constructivists very far. This is no problem for Street, since she's a Humean constructivist. But those who turn to constitutivism hoping to prop up an ambitious moral doctrine will probably need to look elsewhere. Most arguments in this genre focus on agency and action. One can argue in the following way. Adherence to a certain moral principle *M* is constitutive of agency. Practical reasoning presupposes that the reasoner is an agent; this is what makes it practical. Therefore, practical reasoning also presupposes adherence to *M*. So insofar as normative facts are constructed from a practical point of view by practical reasoning, *M* has normative authority for all agents *by default*.

This strategy grounds the authority of the construction procedure in facts that do not themselves admit of constructivist analysis. But it does so in a way that should be palatable to constructivists. The appropriateness of a certain construction procedure is not a brute normative fact and so not something we must be resigned to realism about. Instead, it reflects the conditions on having evaluative attitudes or being an agent and thus, indirectly, the conditions of reasoning practically. These normative facts, the constructivist can say, are nothing over and above facts about *what* valuing, agency, and practical reasoning ultimately *are*. Second, if our concern is to find more "substance" in practical reasoning in hopes of showing that some moral doctrine can be constructed from any practical point of view, a promising place to look is at the metaphysics of the entities that figure in that activity, entities like evaluative attitudes and agency. In particular, if practical reasoning does presuppose agency, then the metaphysics of agency may have implications for practical reasoning that are not immediately obvious.

The hard part of this strategy is showing that adherence to some moral principle is indeed constitutive of agency. There are a handful of arguments to this effect.[4] We'll mention just one.[5] Agency is a natural kind like water or gold. And, like water and gold, there are constitutive requirements on being a member of the kind. But, unlike water and gold, these constitutive requirements are not brute facts about the natural order. Rather, what counts as agency is, in large part, a function of what agents do. This makes agency an "interactive kind": a kind whose nature is partly constituted by what we do. And that, in turn, makes the constitutive requirements of agency rather special. There are no constitutive requirements of agency on a par with having seventy-nine protons or two oxygen atoms. Instead, there is a requirement to behave in such

[4] C. M. Korsgaard, *Self-Constitution* (New York: Oxford University Press, 2009), pp. 177–206 and (in a more complicated way) J. David Velleman, *How We Get Along* (New York: Cambridge University Press, 2009).

[5] Kenneth Walden, "Laws of nature, laws of freedom, and the social construction of normativity", in Russ Shafer-Landau (ed), *Oxford Studies in Metaethics*, vol. 7 (New York: Oxford University Press, 2012).

a way that our behavior (the behavior of me and other would-be agents) constitutes a kind – that our behavior is sufficiently unified, homogeneous, and orderly to qualify as a genuine kind. Not adherence to a fixed standard, but *coordination* with other agents in creating a standard. This coordination takes the form of what Kant calls the "legislation of a Realm of Ends": the making of practical laws that are at once self-given and agreed to by all other rational creatures. Thus, the argument goes, it is a constitutive requirement of agency that we commit ourselves to a certain collective project that turns out to be none other than the creation of a Realm of Ends. If this is correct, then we can understand the appropriateness of a particular construction procedure – the grand construction of value that takes place in the Realm of Ends – as grounded in the demands of agency.

Arguments like this one are premised on the normative significance of agency. But this claim can be resisted. I may wonder why I ought to be an agent rather a kind of creature just like an agent but lacking a key constitutive feature – a "shmagent".[6] That we can entertain this question at all, critics argue, suggests that agency cannot be our Archimedean point.

Given our subject, we will consider a somewhat narrower question: should *constructivists* concede that the shmagency question makes sense? There are two ways to take the challenge. The critic might concede to the constitutivist her account of the constitutive norms of agency, but deny that "agency", so understood, must be all that important to us, much less inescapable. There may be other, equally viable ways of being that we can take up. But the constructivist will want to resist this suggestion, since her goal is to construct *all* of practical normativity from the constitutive norms of agency. The constructivist *cum* constitutivist sees the conditions of agency as the conditions on having a practical point of view at all.

But now the challenge reemerges in a different form. For the critic will likely want to withdraw her initial concession and deny that there are *any* interesting (for example, moral or prudential) norms that are constitutive of *agency* in this broader sense. Indeed, any notion of agency presupposed by practical reason as such ought to be just as amorphous as practical reason as such. So it's not clear why we should expect to extract additional normative content by looking at its constitutive nature. But this objection, it would seem, must be arbitrated on a case-by-case basis.[7]

[6] David Enoch, "Agency, shmagency", *Philosophical Review* 115(2), 2006, pp. 169–98.

[7] Cf. Evan Tiffany, "Why be an agent?" *Australasian Journal of Philosophy* 90(2), 2012, pp. 223–33.

7 Reasoning and other people

We have now surveyed three different approaches to Kantian constructivism. Each of these arguments has a crucial moment when it is suggested that other people play an essential role in an individual's practical reasoning. These moves are pivotal because morality, whatever else it involves, will necessarily include claims about *what we owe to each other*. But we cannot guarantee that all agents will have evaluative attitudes that will ground such obligations by themselves – that everyone will value the welfare of others or take themselves to have reasons to respect their rights. So if we are going to produce a constructivist validation of moral universalism, it will have to be anchored in a claim that the germ of morality can be found in our construction procedure – in practical reason itself.

Each of these moves is also an especially vulnerable part of each respective argument. The argument for the Formula of Universal Law, for example, purports to show that our principles must be universalizable. This standard involves other people insofar as the universalizability of a principle is a matter of whether it can be adopted by those people. But the status that other individuals possess within my practical reasoning because of this requirement seems miniscule. You figure into my practical reasoning not insofar as you have standing to object to what I plan to do (because, e.g., it will harm you) or because you can make demands on me that I must acknowledge. In a sense, you *qua* individual don't matter at all: what matters is whether all persons (a group which happens to include you) could, in principle, adopt my maxim. The concern is that it would be rather surprising if anything worthy of the name "morality" could be constructed from this trifling recognition.

The case we presented for the Formula of Humanity appeals to the demands of anti-parochialism in rejecting the hypothesis that our ends might be systematized by their relationship to the unconditional value of stamp-collecting. Systematicity was supposed to be a demand of reason, but why, one might wonder, think that parochialism is a vice of practical reason? Why think that anyone else's opinion about stamp-collecting is relevant to *my* valuing it?

One part of the argument is particularly prone to this objection. We are imagining a scenario where an agent – call her Clarissa – is investigating the dependence relations of her values with the aim of making them systematically justifiable. The Kantian constructivist says that the best systematizing hypothesis is that Clarissa's own capacity for valuing – her rational nature or "humanity" – is unconditionally valuable. Suppose that Clarissa agrees to all this. But why not stop there? Why must Clarissa allow that the humanity of others is a source of value for their ends, just as her humanity is of

hers? After all, the pressure to make her own value commitments systematically unified doesn't seem relieved by her making any assumptions about anyone else's value. Obviously if the constructivist wants to derive the Formula of Humanity from the process of tracing value-dependency, she needs to show why the objective value of rational nature *as such* better systematizes her ends than this egoist alternative. This project, in turn, seems to turn on whether the systematicity that reason prescribes is intra- or interpersonal, whether reason pushes us to bring just our own ends into systematic coherence or to also accommodate the ends of others. If it's merely intrapersonal, the rationally mandated conclusion would seem to be (merely) that *my* rational nature is unconditionally valuable; *your* value does nothing to explain the value I find in stamp-collecting.

Finally, *pace* the view adumbrated in the previous section, it can be hard to see how the demands of being an agent could introduce an interpersonal dimension to practical reasoning that would ground moral duties. Agency seems to supervene on the agent. Whether I am an agent or a shmagent depends on whether my beliefs and desires are efficacious in the right way, on my having executive control over my actions, on my being free from impairment and coercion, and so on. Robinson Crusoe can be an agent despite his solitude. Other people can certainly interfere with my agency, but there is no condition on my agency that *essentially* makes reference to other people. So it is hard to believe that the demands of agency will be the source of duties or obligations to other people. Or so conventional wisdom says.

Given the importance of these problems, it may be worth confronting the question head on: why should *you* matter to *my* practical reasoning? And not just in the sense of being an object I can use, but in a way that might establish that I *owe you something*.

A potential answer to this question can be found in a striking passage from Kant:

> Reason must subject itself to critique in all its undertakings, and cannot restrict the freedom of critique through any prohibition without damaging itself and drawing upon itself a disadvantageous suspicion. There is nothing so important because of its utility, nothing so holy, that it may be exempted from this searching review and inspection, which knows no respect for persons. On this freedom rests the very existence of reason, which has no dictatorial authority, but whose claim is never anything more than the agreement of free citizens, each of whom must be able to express his reservations, indeed even his veto, without holding back.[1]

There are several intriguing ideas in this passage. The most surprising comes at the end, where Kant says the "claim" of reason consists in the "agreement of free citizens" who are cooperatively engaged in the activity of rational critique. This suggests a radical view about the nature reason: that there is something essentially *social* about the endeavor. This is an important idea, so let's give it a name:

[1] *Critique of Pure Reason*, A738/B766.

> *Sociality of Reason.* Reasoning is (constitutively) a joint activity to which every rational creature is a party.

We won't pause to consider whether this is Kant's considered view.[2] The more pressing question is why we should believe such a thesis. It is, after all, a slightly astonishing thesis. Most of our reasoning, we are probably inclined to think, is solitary, and even when we do reason with other people, it is usually with small and well-defined groups, not all of humanity. So we should be very surprised to learn that an activity was not reasoning simply because it didn't include some of the multitude of rational creatures.

We can only offer the most condensed case for the thesis here. But it starts with observing that, for Kant, reason is a liberating capacity. In the *Conjectural Beginnings of Human History* he explains that most animals are moved by "instinct". For them there is no question of how to respond to an instinctual urge or impulse; a characteristic movement simply follows the instinct. Reason liberates us from this condition by allowing us to "step back" from our own nature and reflect on it: to entertain our instincts as objects of thought subject to interrogation rather than spurs to action. The questions we ask in this interrogation will be characteristically normative ones: given that my instincts are no longer brute forces moving me, I can ask whether I *really* have a reason to act as they would have me act, whether the object they steer me toward *really* is good.[3]

Reason can liberate us from instinct because it is an anti-parochial faculty. It allows and encourages us to step back from our own narrowly animalistic perspective – one where instinct rules – and take up another. It is only from this novel perspective that we can entertain normative questions and, potentially, answer them in ways that produce action contrary to instinct. Reasoning well means subjecting our attitudes, beliefs, and inclinations to the scrutiny afforded by different points of view. Such scrutiny is the difference between *deciding* to act on a reason that I have endorsed and *submitting* myself to the rule of an instinct.

If reason is anti-parochial in this sense, then I cannot simultaneously understand a judgment – that pleasure is good or sodium combustible – as *reasonable* if I understand it as merely *what I happen to think*. For it to be reasonable, I must take it to survive the scrutiny of other points of view, which means, among other things, that I take it to be *justifiable* to those occupying these points of view.

But justifiable how? Onora O'Neill makes one suggestion:

> If thoughts and knowledge claims are to be seen as reasoned, they must at least be followable in thought by others who hold differing views: they

[2] Cf. Kenneth Walden, "Reason and respect", in Russ Shafer-Landau (ed), *Oxford Studies in* Metaethics, vol. 15 (New York: Oxford University Press, 2020) and the papers collected in first part of Onora O'Neill's *Constructions of Reason* (New York: Cambridge University Press, 1989).

[3] See especially 8:111ff. The connection between reflection, freedom, and normativity is recurring

> must be intelligible to those others. If principles of action are to be offered as reasons for action to others with differing ethical and religious commitments, they must at least be principles that could be adopted by those others and used to organize their action.[4]

One could also demand a stronger kind of justification. Perhaps reasoning requires me to convince others to *share* my reasons – to take them as their own – and more generally aims at an ideal of total convergence amongst all agents. This is a difficult question. Fortunately, we don't have to settle it here. Our concern is whether reasoning is constitutively a joint activity. And we seem to have a case for this proposition, whatever standard of justifiability we prefer.[5]

Suppose we are right that the anti-parochialism of reason means that reasoning about a judgment necessarily involves submitting it to the scrutiny of other points of view and, when an actual person occupies one of those points of view, trying to justify it to them. Because reasoning is a holistic business, this justification will end up being reciprocal. You will try to justify your judgments to me, while I do the same to you. And the dyadic case will only be one small part of a massive endeavor, one in which *we* try to justify *our* judgments to *each other* – where "we" includes every creature who can occupy a practical point of view, that is, every rational creature. This suggests that reasoning is a joint activity in which each and every person is a partner.

According to this account, our private episodes of reasoning are best understood as simulations of the real thing. When I am reasoning about whether sodium is combustible or dancing is worth the effort, I am imagining justifying these opinions to various interlocutors who represent particularly salient alternative points of view. I imagine, for example, people who have epistemic access to the chemical properties of sodium or think that the joys of dancing can be replicated by the right sort of video game. According to the Sociality of Reason, this exercise is not reasoning *per se* but a simulation of the reasoning that would go on if we consulted actual persons occupying these points of view. If we are knowledgeable and imaginative, it can be a very good simulation, and since many points of view are not occupied by actual persons at all, we are forced to depend on it. The mistake of many contemporary philosophers is mistaking this simulation of reasoning for the real thing.

This account of reason also offers the possibility for an alternative account of what makes a judgment objective. One conception of objectivity centers on distinctive norms. Suppose there's a chess piece on the table in between us. I say that it's a rook, while you say it's a queen. A few normative claims seem clear here: I have prima facie reason to care about your opinion of the chess piece, at least one of us has gone wrong and

---

theme of *Sources of Normativity*, especially in the second and third lectures.

[4] O. O'Neill, "Constructivism in Rawls and Kant", in Samuel Freeman (ed), *Cambridge Companion to Rawls* (New York: Cambridge University Press, 2003), p. 358.

[5] For a detailed account of reasoning as a social activity, see Anthony Simon Laden, *Reasoning: A Social Picture* (New York: Oxford University Press, 2014).

ought to revise their belief, and convergence on questions about the chess piece is a theoretical ideal for us. By contrast, if you think canary wine is terrific and I think it repulsive, we would be reluctant to say that any of these normative claims follow. This contrast brings out one conception of objectivity. The first sort of judgment is objective, the second isn't. But what explains the difference? Whence the norms that govern "objective" judgments?

The most common explanation locates the difference in the world. There is just one chess piece between us, and it cannot be both a rook and queen. Because theoretical reason aims to accurately represent facts about chess pieces, one of us must have erred. There is nothing analogous to ground the same norms about canary wine. If we take this approach to objectivity in general, then establishing the objectivity of normative judgments becomes a matter of discovering normative entities that can play the same role as the chess piece, and this means establishing the truth of normative realism. The Sociality of Reason offers a different explanation. According to this view, reasoning is, in the first instance, an anti-parochial activity, and the norms that distinguish "objective" judgments are valid *by default* for anything I can subject to reason's scrutiny. They are valid simply because they are constitutive of the process of mutual justification in which reasoning consists. Opinions about canary wine are not objective in this picture because it makes no sense to reason about them: the wine either strikes you as agreeable or not, there's nothing further to ask, no scrutiny to be applied. Normative judgments have a different fate. Their objectivity doesn't turn on normative realism but on whether they merit the scrutiny of reason. And this, in turn, is simply the question of whether there is such a thing as practical reasoning.

What we have here is a pale sketch of an argument for the Sociality of Reason. Reasoning is an essentially anti-parochial activity. It has to be if it is to liberate us from the narrow outlook of our animal nature. What this means, moreover, is that reasoning about a judgment requires subjecting it to scrutiny from other perspectives, and, in particular, justifying it to individuals who occupy those perspectives. This is a joint activity that involves every creature capable of offering and receiving such justifications.

How does recognizing the Sociality of Reason help the constructivist arrive at her Kantian conclusion that every rational agent has reason to be moral? We will mention two possibilities. First, the claim can plug apparent holes the arguments already canvassed. The construction of the Formula of Humanity, for example, depended on the idea that the systematization of one's normative judgments is anti-parochial – that we are aiming not just for systematic values but *systematically justifiable* ends. One can demur from this contention, however, and in doing so the door is opened to rival hypotheses about the ultimate conditions of value. One rival proposes a quasi-Kantian egoism: that the value of *my* rational nature is the unconditioned condition of all value.

Another proposes a quasi-Kantian subjectivism: that for all x, the value of x's rational nature is the unconditioned condition of all value-for-x.[6]

The Sociality of Reason gives us the resources to dismiss these alternatives by justifying and explaining the anti-parochialism of reason. If we are trying to justify our judgments to others, then quasi-Kantian egoism will be an obvious failure. It's not just that no one will agree that I am the ultimate source of all value, but that this claim is so egocentric that it will not be taken seriously. Justifying it to other people would be like trying to justify solipsism to them. Quasi-Kantian subjectivism cannot be dismissed quite so easily, since it's not as baldly parochial. The subjectivist treats her situation as symmetrical to that of her fellow agents: insofar as every x can undertake the kind of reasoning that Clarissa does, x should conclude that x's rational nature is unconditionally valuable *for* x. But the view is still unsatisfying. The subjectivist treats the demands of reason as entirely intrapersonal – as requiring the systematization of an agent's own values – until the very last moment when she acknowledges that there are other agents engaged in reasoning and tries to accommodate this fact by suggesting that all value claims are relativized to individual agents. This is a perfunctory kind of anti-parochialism, analogous to that of the person who first systematizes all her own theoretical judgments about sodium but at the last minute discovers that other people also have perspectives on sodium and tries to accommodate these perspectives in one fell swoop by adopting a simple-minded subjectivism – sodium may be combustible for me but noncombustible for you, water soluble for me but water insoluble for you, and so on. This isn't the utter parochialism of the solipsist, but it's an awkward position.

For one thing, a retreat to this sort of relativism seems like a last resort, at best: to be accepted only if a less relativist alternative cannot be supported. The Kantian constructivist account provides that alternative. For another, the subjectivist story, like the Kantian constructivist one, was supposed to provide justification for our conviction that the things that matter to us *really do matter, normatively.* The story is supposed to offer a supporting *explanation* of their having such value. The subjectivist says our ends matter because we are such that our rational evaluations are value-conferring (albeit only agent-relative-value-conferring). But this seems more like a restating of the phenomenon to be explained than an explanation. The Kantian constructivist story does better on this front. It tells us that our ends are valuable because *we* are valuable – not just valuable *to* someone (as a descriptive, psychological matter) but valuable as ends in ourselves.[7]

We will close by sketching a second, more direct route from accepting the Sociality of Reason to accepting something like Kant's Formula of Humanity. This second route takes very literally the suggestion of that thesis that reasoning is of necessity an interpersonal activity. To see how it goes, let's think a bit more about what would follow

[6] cf. David Sobel, "Is subjectivism incoherent?" *Philosophy and Phenomenological Research* 92(2), 2016, pp. 531–8.

[7] For a longer version of this reply, see Julia Markovits, "Reply to Sobel and Kearns", *Philosophy and Phenomenological Research* 92(2), 2016, pp. 554–5.

from thinking of reasoning as a joint activity. There is more than one sense in which one person may *do something* with another person. As Rae Langton has put it:

> When my friend and I make a cake, I'm doing something with my friend, and I'm doing something with flour, chocolate, cherries, brandy – but there is a difference. My friend, but not the flour, is doing something with me. My friend, and not the flour, is doing what I am doing, sharing the activity.[8]

Langton notes (following Korsgaard) that one place this second, "involved" kind of joint activity is at work is in relationships of mutual accountability, in which the participants hold one another responsible for what they do. Here's Langton again:

> When you hold someone responsible, you are prepared to work with them, view them as someone who has goals of their own that you might come to share, or as someone who might come to share your goals. You are prepared to do something with them, in a sense very different from the sense in which you might do something with a tool.[9]

The basic idea is that properly joint activities constitutively presuppose certain bipolar normative commitments: norms of mutual recognition, accountability, and the observation of rights.[10] If we are to take a walk together, we must recognize each other's rights against being abruptly abandoned. If we are to bake a cake together, we must each recognize the standing of the other's opinions about what kind of cake to bake, whether to use butter or shortening, who should mix and who should measure. And in either of these activities we must be prepared to hold ourselves and each other accountable for violations. These norms are part of what distinguishes a joint activity from one where one person uses another as a tool.

The Sociality of Reason thesis asserts that reasoning is a joint activity. Other rational creatures are our partners in this activity, and so reasoning perforce involves holding oneself accountable to others – justifying oneself to others and acknowledging them as creatures with their own goals, their own rational capacities, and the standing to make claims on us and our deliberations.

So reasoning presupposes versions of the bipolar norms described previously. But reasoning, unlike cake baking, is neither narrowly circumscribed nor something we can opt out of. On the contrary, it is central to everything we do. So the Sociality of Reason entails the universal and categorical validity of these bipolar norms. We're bound to respect all other people's rational capacities, their standing to make claims on us, and

[8] Rae Langton, "Duty and desolation", *Philosophy* 67(262), 1992, p. 487.

[9] Ibid. Cf. Margaret Gilbert, *Joint Commitment* (New York: Oxford University Press, 2016), pp. 23–25.

[10] For more on this concept see Stephen Darwall, "Bipolar obligations", in Russ Shafer-Landau (ed), *Oxford Studies in Metaethics*, vol. 7 (New York: Oxford University Press, 2012).

all their goals *in general*. We, are in other words, bound to always treat them as ends in themselves because they are our necessary partners in reason.

This argument has the advantage of being very direct. It bypasses many of the complications that arise in the other arguments for Kantian ethical doctrines and instead purports to draw one of the doctrines – a version of the Formula of Humanity – directly out of the structural demands of reason. The other side of the coin, naturally, is that it relies on a particularly strong rendering of the idea that reasoning is social. For the Kantian constructivist and her immense ambition, such radicalism may be unavoidable.

23. Constitutivism

On rabbits, hats, and holy grails
David Enoch[3]

[3] For comments on an earlier version, I thank Michael Bukoski, this volume's editors, and especially, Michael Smith.

1 Introduction

If you know – if you *really* know – what a car is, you already know, it seems, what it takes for a car to be a *good* car. For if you know what a car is, you know what its function is (or perhaps what its functions are), how it can be a good instance of the kind *car*, the ways in which a specific car might fail to live up to the ideal-car-paradigm, and the like. So if you really know what a car is, you already know the ways in which cars can be better or worse as cars, the reasons you have to go for a car that satisfies this description or that, and so on. You may not yet know everything about cars – you may lack important information about specific cars (how safe is *this* car? How fuel-efficient? How fun to drive?). But if you know what a car is, you know what would make for a good car, for a car being good *as a car*.

Constitutivism, at a first approximation, is the thought that actions are, in this important respect, like cars. It is the view "that we can derive a substantive account of normative reasons for actions ... from abstract premises about the nature of action and agency" (Smith, 2015, 187). And it is a natural and attractive thought, because it promises to deliver the normativity of practical reasons – and perhaps of morality with it – in a way that is as unproblematic as the sketched view about the goodness of cars.

In particular, if all goes well, a constitutivist metanormative and metaethical theory will deliver the following goods: First, it will allow the right kind of relation of practical reasons to the motivations of those for whom they are reasons, without compromising the objectivity of at least some of these practical reasons (presumably including the moral ones). This is because the motivational features of agents it will tie reasons to will be those that are, in some way, constitutive of agency, and so necessarily shared by all agents. Second, a constitutivist view will be (or at least can be)[1] *metaphysically naturalist*, for there's nothing non-naturalist about the goodness criteria for cars, and it will presumably have all the benefits that come along with a naturalist metaphysics (like perhaps the promise of a simple, respectable epistemology). And third, such a constitutivist view seems to hold great promise as a response to moral and perhaps practical skepticism. If you agree that a certain car is a good car, responding with indifference to this fact seems entirely out of place: One doesn't often hear such things as "But why should I care that it's a good car?" at the car dealership. If constitutivism can show why-be-moral skeptics to be wrongheaded, progress will have been made.

[1] See Smith (2018, 372).

It is not surprising, then, that constitutivism has its ardent supporters, until recently most notably perhaps in Velleman and Korsgaard,[2] and perhaps, if they are right in their drawing on these historical figures, in Aristotle and Kant.[3] Of course, constitutivism has also received its share of critical attention – sometimes in the form of critiques of specific constitutivist views (for different constitutivists fill in the details differently, of course), and sometimes generally.[4] In previous work (Enoch 2006, 2010) I've argued – focusing on Korsgaard and Velleman, but with the explicit aspiration to full generality – that constitutivist views cannot deliver on their promises. Korsgaard and Velleman put forward (different) constitutivist views that attempt to ground normativity in aims or motivations that are constitutive of agency. Korsgaard thinks, roughly, that it's constitutive of action that it be a part of a project in which the agent constitutes herself; Velleman thinks, roughly, that a desire for a special kind of self-understanding is constitutive of agency. And I've argued, roughly, that even if they are right in such claims (and this is a very big "even if"!), still they cannot get normativity out of such constitutive claims, because for anything thus far said, agents have not been given a reason to play the agency-game (rather than, say, the related but distinct shmagency-game). Constitutivists can introduce such a reason, of course, but then this reason itself will not have been accounted for constitutivistically, contrary to the aspiration of giving a constitutivist account of all (practical) normativity.

In recent years, Michael Smith has been developing a distinctive constitutivist view, one that naturally follows in the footsteps of his rightly influential previous work – work that was not obviously constitutivist[5] – and that is different in important ways from Korsgaard's and Velleman's constitutivist views. Furthermore, Smith responds explicitly to my criticism of constitutivism, claiming that at least his version escapes unscathed. In this chapter, then, I focus on Smith's constitutivism.

In the following section, I briefly outline the details of Smith's view. In section 3, I return to the shmagency objection, and I show how, suitably modified, it remains powerful against Smith's version of constitutivism as well. In section 4, I comment on the relation between the shmagency objection and Moorean anti-naturalist thoughts (perhaps captured by the now-notorious Open Question Argument). After offering one more specific objection to Smith's view in section 5, I conclude with some methodological remarks.

---

[2] Korsgaard (2008); Velleman (2000).

[3] And arguably also – quite surprisingly – in Nietzsche. See Katsafanas (2013).

Smith now (2018) thinks that many, many other thinkers are best seen as constitutivist, at least when constitutivism is sufficiently broadly characterized. These thinkers include, according to Smith, David Lewis, Judith Thomson, Jurgen Habermas, Thomas Scanlon, Bernard Williams, Sharon Street, Philippa Foot, and many others.

[4] See here Setiya (2003). See also Fitzpatrick (2005), who to an extent anticipates the Shmagency Objection, to which I am about to get in the text.

[5] Though Smith now seems to claim that his views have always been constitutivist. I get back to this later on.

Again, then, the discussion is rather specific, but, again, the hope is that it will highlight not just problems in a specific (bound-to-be-influential) constitutivist view but rather that it will give at least the feel of general reasons to abandon the constitutivist project as a whole.

2 Smith's constitutivism

The holy grail of moral philosophy, for Smith, is that of securing a robust rational status for morality (2013, 9). And his search for this holy grail goes like this (according to Smith's own helpful summary (though paraphrased) (2013, 25–26)):

> (i) The Dispositional Theory of Value: Final goodness, as indexed to an agent, is fixed by what that agent's ideal counterpart desires.
>
> (ii) Constitutivism, which tells us that helping and not interfering are finally desired by every agent's ideal counterpart.
>
> (iii) The Inheritance Thesis: reasons for finally desiring something inherit their status as reasons from their being considerations that support the truth of the proposition that that thing is finally good.

These three claims together entail that every agent has a reason – indeed, a dominant one – to help and not interfere. Seeing that the reasons to help and not interfere can generate pretty much all of morality (2013, 27–28), morality is safe.

The dispositional theory of value has always been a cornerstone of Smith's moral philosophy (1994), but notice that it alone does not suffice in order to secure the desired rational status for morality. The dispositional theory of value, even coupled with Smith's claims about the relations between reasons and values (here encoded by the Inheritance Thesis), still allows for contingency and variability. In this combination of views, what is ultimately good *as indexed to you*, and so what reasons for action *apply to you*, are functions of the desires of *your* ideal counterpart, and so of *your* desires. And so far, nothing guarantees that the desires of all ideal counterparts will at all converge, or indeed that all – regardless of idiosyncratic desires – will have reasons, say, to keep their promises or not to harm the innocent. Indeed, in earlier work, Smith acknowledged that non-convergence (in the desires of ideal advisors) is a possibility and that if no such convergence is to be had, then metaethical error theory is the way to go (for morality is committed, so Smith argued, to the non-contingency and universality of its content).[1] This is where constitutivism comes in. Smith now thinks that the relevant convergence – the one needed for the holy grail of securing morality's rationality – can be established, because some desires are constitutive of being an ideal agent or advisor.[2] In particular, the (final) desires to help and not interfere are. This

[1] See Enoch (2007, 104), commenting on Smith (2004), and the many references there.

[2] As already noted, Smith now seems to say (2015, 190) that he's always been a constitutivist. It's

way we get both the relation to the reasons, and so to the desires, of the specific agent, because it's *her* ideal counterpart we are talking of, *and* the objectivity needed for morality, because when it comes to *these* desires, all ideal counterparts are necessarily alike.[3]

Why accept, though, constitutivism [as exemplified in (ii)]? Here, Smith relies heavily on Thomson's (2008) idea of *goodness-fixing kinds* (2013, 17 and on). The kind *toaster*, for instance, is a goodness-fixing kind, in that there's something it is to be a good toaster, good *as* a toaster – it's to play the toaster role well or to perform well the *function* of a toaster (to toast bread, make it crunchy without burning it, etc.). There are excellence standards internal to being a toaster. Not all kinds are goodness-fixing (perhaps, for instance, the kinds *storm* and *planet* aren't). But many are – presumably, the opening paragraph shows that *car* is such a kind. And, crucially, argues Smith, so is the kind *agent*. There's something it is to excel, as it were, in the function of agents, or to be a good agent, good *as* an agent, perhaps even maximally good, or ideal, as an agent. The function of an agent is to perform action, and here Smith fills in the details in terms of the standard story of actions as non-deviantly caused by belief-desire pairs. And so, he reaches the conclusion that "A good agent is someone who has and exercises, to a high degree, the capacity to know the world in which he lives and to realize his final desires in it" (2013, 18).

These two capacities, however, cannot guarantee ideal agency, because they cannot guarantee *coherence*. Some kinds of incoherence – presumably, inconsistent with ideal agency – are not so far ruled out. For instance (2013, 21), a desire to believe that p whatever the evidence is problematic because it doesn't cohere with exercising, at a later time, the capacity to know the world (and so to respond appropriately to the evidence). So the ideal agent also has coherence-inducing desires (2013, 22). And these push in the direction of temporal neutrality, so that the ideal agent does not only exercise the capacities to know the world and to realize her desires in it but also cares about her having and exercising these capacities, and they also push in the direction of *personal* neutrality, so that it's not only one's own (present and future) capacities one cares about, but also others'. All ideal agents, then, have the final desires to help and not interfere. And this means that all agents – real, non-ideal agents – have reasons, indeed overriding reasons, to help and not interfere. The holy grail.

---

not clear to me how this can be reconciled with the earlier open-mindedness about the possibility of non-convergence and indeed error theory.

Many years ago, I gave Smith a hard time over what I called the *miracle* of unexplained convergence of all possible ideal advisors that he needed to assume if he were to avoid an error theory (2007, 105–108). I proceeded to briefly discuss possible replies, one of which was to embrace constitutivism (107–108). This, indeed, is how Smith's view seems to have developed. There I also note that going constitutivist will give rise to other problems, like the shmagency worry – in this chapter, I try to make good on that promissory note.

[3] This is the motivation I call, in "Agency, Shmagency" (2006, 34), that of quasi-externalism.

There is a crucial difference between Smith's constitutivism and other constitutivist views like Korsgaard's or Velleman's, and it's important to appreciate it. The desires that ground the rationality of morality, according to Smith, are not constitutive of *action* or of *agency*. It's not that all *agents* have those desires. Rather, all *ideal* agents do. And because of the (purportedly analytic) relation between the desires of ideal agents and the reasons (not-necessarily-ideal) agents have, all (not-necessarily-ideal) agents have reasons to help and not interfere. This is important, because it makes it much easier for Smith to cope with objections that seem devastating for other constitutivist views. Thus, if a feature is constitutive of action or of agency, then nothing that lacks it (or perhaps that lacks it often enough) can amount to action or agency, but then no room in logical space seems to remain for *bad* actions, or *poor* exercises of agency.[4] But this problem does not arise for Smith. What's constitutive of not-necessarily-ideal agency is, according to Smith, only the *function* of agency and so also the relevant idealization: Seeing that you're an agent, then – regardless of your desires and dispositions – the idealization relevant to you, the one sensitive to your function as an agent, is the one that includes also the dominant final desires to help and not interfere.

There is much in this sketched account that I will be granting for the sake of argument.[5] For instance, I will for the most part not question Smith's older dispositional theory of value.[6] Nor will I question the central claim that *agent* is a good-fixing kind, that is, perhaps roughly, that the very concept of agency generates an ordering from best to worst, or perhaps even more roughly, that agency has a function (though for the record, I am *really* not sure about this).[7] Instead, in the following, I will be focusing on the more clearly constitutivist elements of Smith's theory.

[4] On the bad action problem, see, for instance, Lavin (2004), Silverstein (2016, 218–222).

For detailed attempts to defend the inescapability response, see Ferrero (2009, 2018). And for discussion, see Hanisch (2016).

[5] Bukoski (2016) grants Smith less, as indeed he should.

[6] I offer some critical discussion of it in my (2007).

[7] Smith nowhere, as far as I know, offers an argument for this claim, instead characterizing it as something that "should immediately strike us" (2013, 18). See the discussion in section 6.

3 Shmagency's revenge

The initial worry underlying the shmagency objection to constitutivism is perfectly straightforward. That one toaster is better *as a toaster* than another is only normatively relevant – is only something you should care about when about to buy one of the two – if you already have a reason to get *a toaster*, to care about the *constitutive function of toasters*. If you like your bread fresh, and what you're looking for is a toaster-shaped paperweight, or some nice retro piece of kitchen decoration, you may have a reason to get a toaster, but you can perfectly rationally remain entirely indifferent to which of the two is the better toaster. The normative and evaluative relevance of the constitutive features of toasters is entirely parasitic on you already having a reason to care about *it*, or about the *kind* toaster. This seems true in perfect generality – no constitutive condition by itself ever secures normative relevance; its normative relevance has to be grounded in the normative relevance of the kind of which that condition is constitutive. But this creates a problem for constitutivism, because constitutivism, recall, is the hope of generating a *comprehensive* account of (at least practical)[1] normativity in terms of what's constitutive of action and agency. Constitutivists cannot help themselves to any reasons that are not accounted for by the constitutivist story itself. But given the previous lesson, this can't be done – the normative relevance of the standards constitutive of agency has to be grounded in the normative relevance of the kind *agency* or *action*. In other words, for the constitutivist story to get off the ground at all, one has to assume something like a reason to be an agent, or perhaps a reason to care about the function of agency. If agency (toasters) is (are) not normatively relevant to one already, then no irrationality need be involved in failing to satisfy the agency-function (or in getting the toaster that's less good as a toaster). Of course, if we are allowed to assume – before the constitutivist starts telling his story, as it were – that we all necessarily have (dominant) reason to care about *agency*, then the normative significance of the constitutive function of agency is secured but at the price of deserting the hope of offering a fully general account of normativity, for that crucial reason at the ground of the revised constitutivist project has not been given a constitutivist account.[2]

[1] Smith (2015, 195) is even more ambitious than that – he thinks that *all* normativity is accounted for by his account, including epistemic normativity (which is very closely tied, in his account, to practical normativity).

[2] I highlight this feature – how assuming a reason to be an agent undermines the constitutivist aspiration – in "Agency, Shmagency" (2006, 49).

It is in this way, then, that constitutivism may look like it's attempting to pull a rabbit out of a hat (Wiland (2012, 141), also quoted by Smith (2015, 187)). It's attempting to get all the normativity we need or want out of constitutive standards, without relying as input on a normative premise about the normative significance of the kind constitutive of which those standards are. And at this initial level of the worry, it seems to me it applies to Smith's view just as it does to Velleman's and Korsgaard's.

Smith, though, thinks otherwise. Now, I am happy to confirm Smith's suspicion that the somewhat theatrical shmagent-dialogue (2015, 197), which Smith revises to target his own view, is more of a dramatic device than an essential part of the argument. We can safely put it to one side, then, and with it the thought that the challenge is about the constitutivist's obligation to show all of us that we ought to be agents, or some such (a challenge that Smith [2015, 196] rightly rejects, at least vis-à-vis his own view). The challenge, put in more Smithian terms, is to explain why it is the constitutive function of *agency* that is normatively relevant, that in this sense we should care about.[3]

Now, Smith has more to say about the shmagency objection (and I get to some of it shortly), but he nowhere, as far as I could see, answers *this* question, nor does he ever address or even mention the obvious dependence of the normative relevance (as indexed to an agent, perhaps) of the constitutive function of toasters on the independent normative significance of *toasters*. I think that he would respond along the following lines:[4]

> The question why we should care about the function constitutive of agency is a request for reasons, for reasons for caring about those standards; but I've given you *an account* of an agent's reasons, in terms of the function of agency, the function which generates the relevant idealization. To say that you have a reason *just is, as a matter of analytic truth*, to say something about the desires of your ideal counterpart, an ideal itself understood in terms of the function of agency. Reasons, including reasons to care, are *analytically tied* to the function of agency. What more could you want?

This line of response should sound worryingly familiar to you. It is, I think, *precisely* analogous to the so-called analytic justification of induction, and just as problematic. After putting the traditional problem of induction in terms of the question why believe that inductive standards of reasoning are rational or reasonable, Strawson (1952, 248–263) confidently responds that adherence to inductive standards is partly constitutive of reasonableness; a part of what we mean by "rational" or "reasonable" is "in accordance with inductive standards of reasoning". But I find it very hard to believe that anyone at all bothered by the problem of induction is reassured by this kind of response. The worry was never about words ("rational", say, or "reasonable"), and so it's hard to see

[3] This way of explaining the shmagency objection makes it *very* close to Bukoski's (2016) "Normativity Objection".

[4] As Smith was kind enough to confirm in correspondence.

how citing analytic truths can help with it. The worry is about why we should believe the conclusions of inductive arguments, why inductive arguments give reason to believe their conclusions, why inductive standards merit our allegiance, and so on. This is a substantive worry, and no play on words can help coping with it. Note that the point here is not that what Strawson is saying is false (though it may very well be that as well), but that even if it's true, even if rationality is analytically tied to inductive standards of reasoning, pointing this out should not alleviate the worry underlying the problem of induction. Of course, perhaps what Strawson succeeds in showing (and it's not clear to me whether this is what he was hoping to show) is that there's something confused about that underlying worry. Still, if we do take the problem of induction and the worry underlying it seriously, it is very hard to believe that Strawson's analytic "solution" is any solution at all.

The analytic response I put in Smith's mouth is not more satisfying than Strawson's to Hume. To the extent that you feel the pull of the question – why should we care about the function of agency? – you should not be happy with the answer "because this is what 'reason' *means*". As far as I can see, then, the shmagency challenge stands.[5]

Like Strawson, Smith may want to argue that what this shows is that the initial challenge – as presented here – is somehow misguided. Rather than answering the question why we should care about the function of agency, he may want to un-ask the question. But, first, because the question seems to make sense (and obviously does make sense with regard to toasters, say, or cars), this will require considerable argumentative support.[6] And second, if this is the line Smith goes for, it becomes hard to see what the holy grail he was looking for was. The search is presumably motivated by dissatisfaction with a more robustly realist kind of response:

> Why is it rational to be moral, you ask? Why, moral requirements are genuine requirements, they possess full normative power all by themselves, and there's no need for a further story of the relation between morality and rationality. It would be wrong not to help him, and you have a reason to help him, and that's pretty much it.

Dissatisfied with this story, Smith embarks on the quest for his holy grail. But if, at the most crucial part of that quest, he resorts to saying something like "this is just

[5] See also Silverstein's (2016) discussion of the normative inertness of functions and aims.

[6] Perhaps Smith is entitled to rely here on the overall plausibility of his theory, employing the holistic methodology which I discuss in section 6.

Notice that if this is the best available line for Smith, his response to the shmagency objection is at the end of the day surprisingly similar to Velleman's (2009) – both in declaring that the relevant question is unintelligible and in relying on the holistic explanatory value of their respective theories. For my reconstruction of Velleman's response, and for discussion, see my "Shmagency Revisited" (Enoch 2010).

what it means to say you have a reason",[7] then it's very hard to see that progress has been made.

The fact that Smith is interested in securing for all agents not just *a* reason to help and not interfere but *a dominant* one makes his vulnerability to the shmagency objection even clearer. The reasons to help and not interfere are, if everything else in Smith's account works, grounded in the desires constitutive of ideal agency. But all of us also have many other reasons. Yours, for instance, are grounded in the more idiosyncratic desires that your ideal counterpart has, in virtue of the desires that *you* have. What should we say, then, about cases in which these different reasons are in conflict? Think toasters again. Suppose that some toasters are also beautiful, or that they make a lovely sound when the toast is ready. It seems farfetched to think that this makes them better *as toasters*, but it seems very natural to think that it makes them *better*, and in particular, that, say, among equally good toasters (as toasters), it can be perfectly rational to prefer the one that's more aesthetically pleasing. Furthermore, there need be nothing irrational about preferring the more beautiful or better-sounding toaster to another one that is somewhat better at playing the toaster-role, or at fulfilling the function constitutive of the kind toaster. All of these are just more reasons, more considerations that count in favor of one toaster or another. That some of them are closely related to the toaster-function seems clearly normatively irrelevant (as is evidenced, for instance, by the silliness of the remark "Sure, that it makes a lovely sound makes it better, but does it make it better *as a toaster*?" Who could possibly care about *that*, at least in the context of choosing a toaster?). Back to actions, then: Suppose I have a reason to help my neighbor who is in need. I also have many other reasons, reasons that are also grounded in the desires of my ideal agent, except in the more idiosyncratic ones, like perhaps the reason to have a lot of rest (philosophy is very hard work). Why think that – in a case in which I can either help my neighbor or get some rest but not both – the former gets normative priority over the latter? The point is not just that such priority is something Smith explicitly commits himself to (when talking about the *dominant* desires of ideal agents to help and not interfere). The point is that he *has to*, on pain of not securing the robust foundation for morality that he is after. Why think, though, that the reasons that are closely related to the function of agency have such priority over all other reasons? The mere fact that *they are* related to the function of agency (as I'm here granting for

[7] In fact, and as Smith explained in correspondence, things are more complicated than this. Smith believes that what is analytic is the connection between being of value (for one) and being desired by one's ideal advisor. If someone's thinking is not disciplined by this kind of connection, claims Smith, we can't see them as thinking about value at all. The claim that what's of value (for each) is to help and not interfere follows from this analytic claim, and is therefore necessary and a priori, but is not quite analytic in the same sense (we can recognize someone as thinking of value even if their thought is not disciplined by these further claims).

The point in the text has to be reworded in order to accommodate these complications. But once suitably reworded, it will retain, I think, its force.

the sake of argument) seems normatively irrelevant, just like the parallel fact in the toaster case. And this – that being constitutive of agency is normatively irrelevant – was precisely the underlying shmagency worry.[8]

[8] In the context of arguing against the shmagency objection, Smith (2015, 195) addresses also Tiffany's (2012) related discussion. Let me note that I don't see how what Smith says in response to Tiffany deals with the best understanding of his challenge. This is the dilemma – anticipated in "Agency, Shmagency" (41) – of either packing so much into the relevant constitutive conditions so that it's implausible that they *are* constitutive of action, agency, or ideal agency; or of working with a thin enough notion of agency to make these claims plausible, at the price of making it hopeless to derive all of morality from what's constitutive (in this sense) of agency.

4 Is it just the open question argument all over again?

Following Wiland (2012, 138), Smith (2015, 198) raises the suspicion that at the end of the day, what's really going on with the shmagency objection is just a rehearsal of Moore's infamous Open Question Argument (OQA), or something in its very close vicinity.[1] Smith meets this possibility with impatience: "If this is the right diagnosis of Enoch's hostility toward constitutivism, then it seems to me that we should simply take note of his Moorean intuitions and move on" (2015, 198).[2] It won't come as a huge surprise that I take such Moorean intuitions more seriously than Smith does,[3] and I agree that the shmagency worry is related to such general anti-naturalist intuitions. But it's not *merely* a retelling of that same story. So it may be worth our while to say more on this relation. I want to make two points.

First, Moorean, OQA-like thoughts are very general. Once applied to a particular naturalist theory, they can be seen not just as a reason to reject it (just like any other naturalist theory) but also as generating a prediction that a more specific problem will arise: If such Moorean thoughts are correct, all naturalist theories are bound to fail, because they attempt the impossible (bridging the natural-normative gap, or some such). But they can fail in interesting, subtle, and *different* ways. If Moore is right, we know, for each naturalist theory, that it hides an illegitimate move from the natural to the normative *somewhere*, but this doesn't mean we know *where*. The shmagency objection finds this place, if I'm right, for all constitutivist theories. As applied to Smith's version of constitutivism: because Smith's is a naturalist theory, we've known all along that an illegitimate step is there somewhere, and the shmagency objection shows that it's in the implicit assumption (roughly) that the constitutive function of agency is fundamentally and intrinsically normatively relevant (for instance, in a way that the constitutive function of toasters is not). Once we see that, we have not only a refutation of Smith's theory but also a (modest) confirmation of the general Moorean prediction, that something of this kind goes wrong with any naturalist theory. It's the

[1] Wiland merely notes some affinity, Smith's suspicion is more, well, of a suspicion. For an extended discussion of constitutivism and the is-ought gap (mostly not in the context of discussing the shmagency objection), see Silverstein (2016).

[2] See here also Rosati (2016, section 8.3.1). What follows in the text may be seen also as a partial response to Rosati.

[3] See my discussion in *Taking Morality Seriously* (Enoch 2011, 100–109).

same with rabbits and hats, really: it's one thing to know there's a trick *somewhere*, it's another to find it.

Second, constitutivists, I take it, never thought of themselves as *plain-old* naturalists. They were offering, I think, a *really special* kind of naturalist reduction of morality and of normativity, one that was somehow better placed, compared to straightforward naturalist reductions (à la Cornell realists, for instance) to deal with Moorean worries and intuitions.[4] The hope was, I think, that by relying on what is constitutive of agency, or perhaps on the function of agency and so on what is constitutive of ideal agency, we would secure the kind of objectivity that any objectivist naturalist reduction can guarantee but without the objectionable externalism or alienation that usually come with it, because the right kind of connection to the relevant agent's desires or reasons will have been secured. At least a part of the hope, I think, was that a constitutivist reduction would close questions that on other naturalist reductions remain open. And if constitutivism does not enjoy such an advantage over other naturalist reductions, it is not at all clear that it is overall more plausible than these other views are. The shmagency objection shows, then, that constitutivists are not better positioned than other naturalists vis-à-vis Moorean worries. Indeed, they may be in a *worse* position than others, for constitutivism accords utmost significance to the function of agents (as generating real reasons), but not to the functions of other goodness-fixing kinds, like toasters. But at least so far, Smith hasn't explained – nor has anyone else, as far as I can tell – why it is that agency is so unique among goodness-fixing kinds. This is a burden that the constitutivist naturalist reduction faces and that other naturalist reductions may not need to face. That constitutivists are not better positioned vis-à-vis Moorean worries, and that they may be worse positioned, compared to other naturalists is, given the dialectical situation, not an insignificant result.

[4] I think this line of thought is present in Rosati (2003, 2016), throughout, though perhaps from a critical distance.

5 Equivocations

Many of the propositions Smith asserts are, considered on their own, highly plausible, yet when put together, leave one (well, me) with the feeling that some rabbit was just pulled out of a hat. This raises a thorny dialectical issue: Viewed in one way, this is precisely the way to make progress in philosophy – by noticing surprising inferential relations between initially plausible propositions and using them in inferences to surprising conclusions. Viewed in another way, though, this is dangerous, and the name of the danger is "equivocation". If, for instance, Smith says one plausible thing about the relation between coherence and ideal agency and then another plausible thing about the relation between ideal agency and what one has reason to desire, and if put together, these claims entail highly surprising claims about what one has reason to desire; then this is some reason to suspect that Smith has subtly equivocated on "ideal agency". Admittedly, things are tricky here, because – as I noted already many years ago commenting on Smith's work (my 2007, 100) – it's not as if whenever one is putting forward an argument one is also required to argue for the claim that one is not equivocating in that argument. Still, sometimes we can do more than merely point out the general equivocation worry. Sometimes we can support more specific accusations of equivocation, and when we can, if you still want to defend the attacked argument, you should step up to the plate and offer reasons to reject the equivocation charge. I want to offer some reasons to think Smith's argument is guilty of such equivocation and also to offer a general equivocation-detection test and claim that Smith doesn't pass it.

Consider, then, the following (2013, 18–19): "These are the final desires that an agent should have, *in the sense that* his having those final desires is required for him to meet the highest standards that are internal to the concept of agency." (my italics). I am happy to grant Smith this claim. Notice that Smith himself felt the need to add here the "in the sense that" clause, apparently acknowledging that there are other senses, and that the unqualified sentence ("These are the final desires that an agent should have.") may be false. But this caution disappears later on, when Smith takes himself to have established that these are, well, the desires that an agent should have, or has most reason to have, period. But this does not follow, of course. In particular, it doesn't follow that failing to have these desires is irrational, or that if morality is tied to these desires, then its status is secured. All that follows is that if morality is tied to these desires then it is rational *in the sense that*... . Dropping the qualification results in a fallacy.

Or consider what Smith has to say about the relevant notions of *ideal agency* and *coherence*. He says such things as "it is a contradiction in terms to suppose that an ideal agent's psychology is not maximally coherent" (2013, 21), and we can happily grant as much, for *one* sense of "ideal" and for *one* sense of "coherent". Perhaps, for instance, a sufficiently thin notion of coherence is indeed a necessary condition for ideal agency, in the sense of "ideal" in which (perhaps) one always has a reason to desire what one's ideal counterpart does in fact desire. Perhaps. But Smith also thinks that the desire "to not now interfere with her exercise of her capacity to have knowledge of the world in which she lives" is "plausibly thought to be constitutive of being ideal" (23). And now one gets a glimpse of rabbit ears: Is this desire plausibly considered a necessary condition for coherence, say, *in the same sense* in which incoherence is always a rational flaw? Is this understanding of an ideal agent the very same understanding that makes it plausible that one always has reason to act in the way that one's ideal counterpart desires that one act? Perhaps this surprising claim can be supported, but surely it is not supported merely by a tendentious use of the words "ideal" or "coherent".

Bukoski (2016, 123) notes a closely related ambiguity in Smith's use of "ideal agent": he distinguishes between what he calls a *kind-ideal agent* (fixed by what's constitutive of agency) and a *rational-ideal agent* (fully adhering to plausible standards of rationality). Bukoski doesn't use the word "equivocation", but in effect he criticizes Smith for moving back and forth between these two kinds of ideal agency.[1] Much of what Smith says about ideal agents sounds plausible, Bukoski in effect says, because we choose the notion of ideal agency that renders it plausible, but for the soundness of Smith's arguments, what's needed is that all these claims be plausible *about the same kind of ideal agency*, and this just isn't so.

Here is a suggestion for a general test for equivocation. If a term is introduced, and if the plausibility of some of what's being said (like that of one of the premises that are put using the term) crucially depends on the term used to say it, then equivocation is probably present. In the other direction – if dropping a term, and exchanging it with an explanation of its meaning, renders an argument less convincing, then the initial argument owed at least some of its plausibility to an equivocation on the relevant term. After all, if the claims made really are plausible, and don't just *sound* plausible, then their plausibility should survive rephrasing.

I don't think that Smith's arguments pass this test. This, I think, is a way of strengthening Bukoski's point. After all, there's no magic in the *term* "ideal agent". So if Smith's arguments work, they should work also when we replace all occurrences of "ideal agency" with their explanation. But they do not: The claim that actual agents have reason to do what their ideal counterparts desire is perhaps plausible under one understanding of "ideal agent", and the claim that all ideal agents will have dominant desires to help and not interfere is plausible under some understanding of "ideal agent", but it's not the case that both are plausible under *the same* understanding of "ideal

[1] Bukoski was kind enough to confirm, in email correspondence, this reading of him.

agent". Drop the words "ideal agent", and see whether you can tell a remotely plausible story here. If, as I predict, you cannot, this shows that Smith's argument owes much of its seeming plausibility to the attractiveness of the *term* "ideal agent". And this means he's in all likelihood guilty of equivocation.

Here's another way of seeing this: Recall the toaster intuition, according to which *agent*, like *toaster*, is a goodness-fixing kind, so that there are standards internal to agency (perhaps its function) that dictate what it is to be a good agent. If the plausibility of Smith's arguments is not to depend on suspicious uses of the term "ideal", then it must be the substantive, underlying toaster intuition that carries over to everything Smith wants to say about ideal agents. But how plausible is it that an agent who fails, say, to have a dominant desire to help distant strangers fails as an agent, *in the same sense in which a toaster that routinely burns slices of bread* fails as a toaster? By the time Smith gets to this part of his discussion, he seems to have forgotten the intuition underlying talk of ideal agency (kind-ideal, in Bukoski's terms) and to just continue with the *word* "ideal" and with connotations that come along with it regardless of the more precise sense given to it in the theory. In other words, he is equivocating on "ideal".

Perhaps Smith is entitled to respond, though, by highlighting the overall explanatory payoffs of his theory. If his theory is very good in other respects – certainly, if it's the best theory of normative thought and practice overall – then perhaps we should accept that there is no equivocation going on, on the strength of the plausibility of the theory as a whole. So we need to (briefly) talk methodology.

6 Methodology

Smith emphasizes that his theory, like all others, will eventually have to be judged both comparatively and holistically. He claims tremendous explanatory advantages for his theory – not just the holy grail of moral philosophy, but also the right relations to the philosophy of action, a plausible account of moral responsibility (of which I've said nothing here), and more. And he concludes (2015, 199): "Whether a theory like Enoch's can provide a similarly unified and parsimonious theory of reasons for belief, desire, action, and responsibility is yet to be seen. My hunch is that it cannot."

I fully agree that this is the right methodology here (and in philosophy in general). Competing theories are to be evaluated holistically and comparatively. This means that a theory's (theoretical) advantages can sometimes compensate for its shortcomings. I find the metaphor of *plausibility points* helpful here: Each theoretical advantage earns the theory plausibility points, for each disadvantage it loses such points, and at the end of the day, we should go for the theory that has the most plausibility points overall. This means, for instance, that even if the analytic response to the shmagency objection (or the related way of denying the intelligibility of the challenge) initially seem implausible, it could be rendered acceptable if it follows from our best theory overall on the strength of its other many advantages. Smith and I are, then, on the same page when it comes to philosophical methodology. Still, I need to make here two further points.

First, though the final determination of which theory to accept is going to have to be comparative, it doesn't have to be *personal* in the way the quoted sentence suggests. After all, it's not about *me*. It's quite possible, then, that the shmagency objection (or perhaps the equivocation one from the previous section) defeats Smith's theory – or anyway, makes it lose too many plausibility points – but that *my* positive theory does even worse and that the theory that gets the highest plausibility score ends up being another one altogether. So even if Smith's hunch is right, and my Robust Realism doesn't get an overall higher score than his theory, this doesn't save him from my objections. Relatedly, the overall plausibility of Smith's theory will be determined not just by the strength of "my" objections or the objections in *this* chapter. Recall that I've granted Smith quite a bit for the sake of argument. Also, I haven't even hinted at a discussion of the normative implications Smith claims for his theory – a point in which Smith thinks he gains plausibility points,[1] but, as Bukoski has convincingly

[1] At one point (2013, 26) Smith claims that the normative judgments that flow from his theory are *strikingly* similar to those of commonsense morality, apparently taking this striking, true prediction

argued (2016, 137–141), he loses quite a few. So things may look even worse for Smith's theory than (the rest of) this chapter suggests.

Second, the fact that the final determination (of which theory to accept) will be holistic doesn't mean that there's no room for local argumentation. We need to be clear about more local arguments, if only in order to get the right plausibility-point-input into the final stage of tallying them. The equivocation suspicion from the previous section should cost Smith rather heavily in plausibility points, and this is important even though it's not impossible that his theory's many other advantages will make us decide that there was no equivocation after all. And if Smith loses plausibility points over the shmagency objection – certainly, if this is true of *any* constitutivist theory – then this is an important result, even if we keep an open mind about the possibility of this loss being offset by other gains. For now, I am happy to settle for this result.

to confirm his theory. Now, if Bukoski is right in arguing (2016, 137–140) that Smith's theory entails *implausible* first-order judgments (and for the record, I think he is), then really the theory is *dis*-confirmed. But even if this is not so, one gets the feeling that much of the theory was tailored to fit morality (this is clearest about the extension of the relevant notion of coherence from desires about one's present and future self to desires about others). So Smith's theory doesn't *predict* morality, but at most *accommodates* it, and in a way that does not earn it more plausibility points. The striking similarity, in other words, was a requirement for the development of the theory, and so is not that striking after all.

7 Conclusion

Is constitutivism, then, without hope?

Let me remind you that in this chapter, the discussion – even if entirely successful – defeats only Smith's version of constitutivism. It is not impossible that other versions can do better – though I have to say that most of the worries seem general to me, sufficiently general to give rise to the suspicion that they (or their close cousins) will arise for any other equally ambitious constitutivist theory. Still, philosophy is hard, and it's not impossible that a version of constitutivism will be put forward that is – even if not immune to the worries here and related worries – better positioned to address them.

But perhaps the way forward for constitutivism – as for so many other theories, in philosophy and elsewhere – proceeds via reducing expectations. If constitutivists are willing to think of their constitutivism not as an attempt at a perfectly general and complete metaethical and indeed metanormative account but rather as a partial account of the relevant normative subject, one that relies on, say, a Moorean or Platonic reason to be an agent, or on a naturalist reduction in other terms, perhaps the main insights of constitutivism can still be maintained. It remains to be seen whether this compromise position is one constitutivists will find acceptable.

References

Michael Bukoski (2016), "A Critique of Smith's Constitutivism", *Ethics* 127, 116–146.

David Enoch (2006), "Agency, Shmagency: Why Normativity Won't Come from What Is Constitutive of Agency", *Philosophical Review* 115, 169–198.

David Enoch (2007), "Rationality, Coherence, Convergence: A Critical Comment on Michael Smith's *Ethics and the A Priori*", *Philosophical Books* 48, 99–108.

David Enoch (2010), "Shmagency Revisited", in Michael Brady (ed.), *New Waves in Metaethics* (Basingstoke: Palgrave), 208–233.

David Enoch (2011), *Taking Morality Seriously: A Defense of Robust Realism* (Oxford: Oxford University Press).

Luca Ferrero (2009), "Constitutivism and the Inescapability of Agency", *Oxford Studies in Metaethics* 4, 303–333.

Luca Ferrero (2018), "Inescapability Revisited", *Manuscrito: Revista Internacional de Filosofía* 41 (4): 113–158.

William Fitzpatrick (2005), "The Practical Turn in Ethical Theory: Korsgaard's Constructivism, Realism and the Nature of Normativity", *Ethics* 115, 651–691.

Christoph Hanisch (2016), "Constitutivism and Inescapability: A Disagnosis", *Philosophia* 44, 1145–1164.

Paul Katsafanas (2013), Agency and the Foundation of Ethics: Nietzschean Constitutivism *(Oxford: Oxford University Press).*

Christine M. Korsgaard (2008), The Constitution of Agency: Essays on Practical Reason and Moral Psychology *(Oxford: Oxford University Press).*

Douglas Lavin (2004), "Practical Reason and the Possibility of Error", *Ethics* 114(3), 424–457.

Connie S. Rosati (2003), "Agency and the Open Question Argument", *Ethics* 113, 490–527.

Connie S. Rosati (2016), "Agents and 'Shmagents': An Essay on Agency and Normativity", *Oxford Studies in Metaethics* 11, 182–213.

Kieren Setiya (2003), "Explaining Action", *Philosophical Review* 112, 339–393.

Matthew Silverstein (2016), "Teleology and Normativity", *Oxford Studies in Metaethics* 11, 214–240.

Michael Smith (1994), *The Moral Problem* (Oxford: Wiley-Blackwell).

Michael Smith (2004), *Ethics and the a Priori* (Cambridge: Cambridge University Press).

Michael Smith (2013), "A Constitutivist Theory of Reasons: Its Promise and Parts", *Law, Ethic and Philosophy* 1, 9–30.
Michael Smith (2015), "The Magic of Constitutivism", *American Philosophical Quarterly* 52, 187–200.
Michael Smith (2018), "Constitutivism", in *Routledge Handbook of Metaethics*, edited by Tristram McPherson and David Plunkett (London: Routledge), pp. 371–384.
P. F. Strawson (1952), *Introduction to Logical Theory* (London: Methuen).
Evan Tiffany (2012), "Why Be an Agent?" *Australasian Journal of Philosophy* 90(2), 223–233.
Judith Jarvis Thomson (2008), *Normativity* (Chicago: Open Court).
David Velleman (2000), *The Possibility of Practical Reason* (Oxford: Oxford University Press).
Eric Wiland (2012), *Reasons* (London: Continuum).

24. Reasoning first

Pamela Hieronymi

Philosophers have been fascinated, lately, with reasons, as such. Though we boast a long history of attention to Reason, and even to reasoning, the current interest is in reasons – though not in particular reasons but in reasons as a class, reasons *per se*. We would do well to consider how we were led to this interest and what we hope to gain from it. I recommend both questions for consideration, but, for better or worse, I will not pursue them here. My own view is that we were led here, in the main, by skeptical concerns, and what we have to gain, in the main, is a better understanding of our agency, together with the avoidance of important confusions. Unlike some, I doubt that the study of reasons, as such (or, for that matter, the study of Reason or of reasoning), will tell us very much about what we ought to do or about what is good or of value.

However, rather than defend these positions here, I will instead argue that we should think about reasons, as such, differently than many have been thinking of them. Many think of reasons as facts, propositions, or considerations that stand in some relation to attitudes, actions, events, states of affairs, or perhaps some other fact or consideration. They think of the relation as either an explanatory one or, as they put it, a "normative" one. I will suggest, instead, that we should see reasons as items in pieces of reasoning. Reasons relate, in the first instance, not to psychological states, events, or states of affairs, nor even to other considerations, but rather to questions. Their relation to a question is neither explanatory nor "normative." If we must give it a label, we could call it "rational" – but the label will be uninformative: it would mean only that the reason bears or is taken to bear on the question.

After presenting (my understanding of) the current way of thinking about reasons, I will sketch three difficulties that arise when you think of reasons in this way. The chief benefit of the alternative is that, by relating reasons first to questions, we bring rational agency into view: It is the thinker, the rational agent, who settles questions and therein forms attitudes or sets themselves to act – that is, to bring about events or states of affairs. The thinker thus mediates between considerations, on the one hand, and attitudes, actions, events, or states of affairs, on the other. In contrast, the current way of thinking, which relates reasons directly to attitudes and actions, occludes rational agency – it hides the use of reasons in thought. By bringing the thinker into view, the alternative can avoid the difficulties that arise on the current way of thinking. (This chapter draws on two previous papers, which started life as a single paper arguing for this position – that we should think of reasons as items in pieces of reasoning.[1] I here complete the original ambition.)

[1] Pamela Hieronymi, "Reasons for Action," *Proceedings of the Aristotelian Society* 111 (2011); "The Wrong Kind of Reason," *The Journal of Philosophy* 102, no. 9 (2005).

Current thinking

To start, consider the current thinking. It is common for philosophers, when thinking about reasons, to begin with the thought that reasons explain. The fallen tree explains the power outage, and it is the reason for the outage. The fact it is an El Niño year explains the heavy rainfall and is the reason for the rainfall. Thought of in this way, it is easy to conflate reasons and causes – though we would do well to remember that causal relations and explanatory relations differ. As P. F. Strawson puts the point:

> if causality is a relation which holds in the natural world, explanation is a different matter. People explain things to themselves or others and their doing so is something that happens in nature. But we also speak of one thing explaining, or being the explanation of, another thing, as if the explaining was a relation between the things. And so it is. But it is not a natural relation in the sense in which perhaps we think of causation as a natural relation. It does not hold between things in the natural world, things to which we can assign places or times in nature. It holds between facts or truths.[1]

In addition to explaining, reasons can play very different roles: they can justify, or count in favor, or show correct, or be grounds for. The fact that it is nearly dinnertime counts in favor of leaving the office and is a reason for leaving. The fact that she was innocently unaware of the problem is the reason for her silence and justifies her silence. Justifying, counting in favor, and showing correct have been lumped together, recently, under the label "normative." One now-standard approach analyzes these so-called "normative" relations as multi-place relations. For example, T. M. Scanlon claims that a reason in what he calls "the standard normative sense" is a four-place relation, holding between a fact, a person, a circumstance, and an action or attitude of that person.[2] Referring to the same class, John Skorupski adds further variables and complexity.[3]

There seems to be broad agreement on the basic divide into explanatory and "normative."[4] Once we move beyond it, matters become more controversial, and some of the underlying difficulties start to appear. I will consider three.

---

[1] Peter F. Strawson, *Analysis and Metaphysics* (Oxford: Oxford University Press, 1992), 109.

[2] T. M. Scanlon, *Being Realistic About Reasons* (Oxford: Oxford University Press, 2014), lecture 2.

[3] John Skorupski, *The Domain of Reasons* (Oxford: Oxford University Press, 2010), chapter 2.

[4] A noteworthy view that does not start with this divide, and that is, I think, compatible with the position I advance here, is John F. Horty, *Reasons as Defaults* (New York: Oxford University Press,

2012).

The first difficulty

The first such difficulty is that the counting-in-favor-of, or "normative," relation can seem more mysterious than the explanatory relation; in fact, it can seem itself to require explanation. Thus, philosophers sometimes take the explanatory relation as primitive and claim that the "normative" relation holds when a consideration explains something – often something about value (broadly speaking). For example, John Broome identifies 'normative' reasons (specifically, 'perfect' reasons) as facts that explain 'ought' claims.[1] For Jonathan Dancy, reasons are grounded in values.[2] Daniel Fogal argues that reasons are considerations that explain what there is reason to do – where "what there is reason to do" is not understood in terms of reasons but rather in terms of "normative support."[3]

Other philosophers, still taking the explanatory relation as primitive, characterize "normative" reasons as those that explain something "non-normative." For example, in his early book, Mark Schroeder claimed that a consideration is a "normative reason" for action if (roughly) it is part of an explanation of why a given action would satisfy some desire.[4] Stephen Finlay understands "normative" reasons, generally, as explanations of why something is good, and then he gives reductive, "end-relative" account of good.[5]

By explaining reasons for action in terms of something else, these philosophers risk an additional sort of worry. The worry is brought out clearly by Schroeder, who notes that his own view – according to which what explains your reason for action is, in every case, the possible satisfaction of some desire of yours – may make acting for reasons seem "objectionably self-regarding." By explaining the reason by appeal to desire-satisfaction, it seems that Schroeder has turned us all into a certain kind of hedonist or egoist.

Schroeder points out, though, that the objectionably self-regarding objection depends on what he calls the "no background conditions" view, which holds that any consideration that explains why some other consideration is a reason for action itself becomes *part* of the reason for acting. Schroeder denies this. He believes the facts that

[1] John Broome, "Reasons," in *Reason and Value: Themes from the Moral Philosophy of Joseph Raz*, eds. R. Jay Wallace, et al. (Oxford: Clarendon Press, 2004).

[2] Jonathan Dancy, *Practical Reality* (Oxford: Oxford University Press, 2000), 29. Dancy does not explicitly say that value explains reasons, only that value grounds reasons. I do not know how he understands grounding and explanation.

[3] Daniel Fogal, "Reason, Reasons, and Context," in *Weighing Reasons*, eds. Errol Lord and Barry Maguire (New York: Oxford University Press, 2016). Pg. 13.

[4] Mark Schroeder, *Slaves of the Passions* (Oxford: Oxford University Press, 2008), 224.

[5] Stephen Finlay, *Confusion of Tongues: A Theory of Normative Language* (New York: Oxford University Press, 2014).

explain why a consideration is a reason for action stay in the background. Because facts about your desires do not become part of your reason for acting, your action does not become objectionably self regarding.[6]

However, Schroeder points out, with puzzlement, that most philosophers assume the no-background-conditions view. He thinks this strange, noting that the facts that explain a thing do not typically become part of that thing: the fact that someone is elected and inaugurated explains why that person is president, but those facts do not become part of the president.[7]

I side with the majority, here, thinking that considerations which explain why a fact is a reason for action typically become "part" of the reason to act. For support, I would first point to the intuitiveness of a collection of problems that would not otherwise arise. These include not only the self-regarding objection Schroeder hopes to avoid, but also the rule-worship objection to rule utilitarianism and a handful of concerns that moral theory provides the "wrong" or "ulterior" motives for moral action. Pritchard famously thought that moral philosophy rests on a mistake because, in trying to explain why you must do your duty, it provides an ulterior motive.[8] Kant, before him, accused all previous moral theories of making the same error.[9] Williams worried that, by justifying saving your spouse, moral theory would (peeve your spouse and) attack your integrity, alienating you from your own motives.[10] Along the same lines, one might worry that reflective and thoughtful divine command theorists can only practice piety, while Kantians can be only conscientious (or concerned with coherence) – one can worry that explanations spoil virtue. I refer to this collection of concerns as the problem of "bleed through" – and it depends on the no-background-conditions view: it depends on the thought that the explanation of why a reason for action is a reason for action will, if believed, become part of one's reason for acting and so color one's motivations. While Schroeder would point to this collection of problems to support his claim that philosophers are widely committed to the no-background-conditions view (and he seems to think we can avoid these problems by denying it), I would instead point to the intuitive appeal of these problems as support for no-background-conditions

[6] Schroeder, Chapter 2.

[7] Schroeder, 24. Following Amber Kavka-Warren, I will note that the illustration is off. If reasons are considerations, or facts, then at issue is not whether the fact that someone was elected and inaugurated is part of the *president*, but rather whether it is part of *the fact that* that person is the president. While it is clear that it is no part of the human, it less clear that the one fact is not in some way "part" of the other. Talk of "parts" is unclear, in this context.

[8] H. A. Pritchard, "Does Moral Philosophy Rest on a Mistake?" *Mind*, 21, n. 81 (1912).

[9] Immanuel Kant, *Groundwork of the Metaphysics of Morals*, 4: 444.

[10] Bernard Williams, "Persons, Character, and Morality," in *Moral Luck* (Cambridge: Cambridge University Press, 1981). See also "Internal and External Reasons," in *Moral Luck* (Cambridge: Cambridge University Press, 1981); "A Critique of Utilitarianism," in *Utilitarianism: For and Against*, eds. J. J. C. Smart and Bernard Williams (Cambridge: Cambridge University Press, 1973). For an interpretation relating these three papers, see Pamela Hieronymi, "Internal Reasons and the Integrity of Blame," [Unpublished Manuscript, [[https://ucla.box.com][https://ucla.box.com/v/HieronymiIntegrityOfBlame]].

view. They seem genuine problems, problems that do not simply disappear with the assertion that explanations stay in the background. I will later be in a position to say a tiny something about why.

For now, we are cataloging difficulties. The first difficulty is the apparent mystery of the counting-in-favor-of relation. Some explain it terms of ought facts or value, others would reduce it to some "non-normative" relation, and at least one person, Scanlon, simply takes it as primitive.[11]

[11] T. M. Scanlon, *What We Owe to Each Other* (Cambridge: Harvard University Press, 1998), Chapter 1.

The second difficulty

A second difficulty appears when we consider, not the explanation of the counting-in-favor-of relation, but rather the explanation of actions done for reasons. It is sometimes thought that reasons for action explain action by providing motivation to act. The fact that I am hungry not only explains my eating, but it also motivates me to eat, and, one might think, it is my reason for eating. Likewise, the fact that she betrayed me motivates me to avoid her, and it is my reason for avoiding her.

Once we note that we can explain actions that are themselves done for reasons, we may want to ask the question with which Donald Davidson opened "Actions, Reasons, and Causes:" "What is the relation between a reason and an action when the reason explains the action by giving the agent's reason for doing what he did?"[1] That is, we may hope to understand the role of the reason for which the person acted – the agent's reason, as Davidson calls it – in the explanation of the action.[2]

The most simple of views would explain the action simply by appeal to the agent's reason: If Jae left the store because it was closing, then the store's closing – Jae's reason for leaving – explains Jae's departure.

Difficulties for the simple view appear when we remember that people are fallible: Perhaps Jae was mistaken; she thought the store was closing, but it was not. Fallibility generates two types of difficulty.[3]

First, if the store was not closing, then there was no reason to leave. Yet Jae's action was undertaken for a reason – she did not act on a whim, for no reason. It seems we must say that she acted for a reason that was no reason.[4] To make sense of this, we need a way to talk about the considerations that someone took to count in favor of an action, on the basis of which they acted, whether or not the considerations actually counted in favor of acting. Scanlon calls these "operative" reasons; many others call them "motivating" reasons.[5] Davidson has something like this in mind when referring to "the agent's reason for doing what he did." So, it seems, this worry might be met by making a distinction and introducing new terms.

---

[1] Donald Davidson, "Actions, Reasons, and Causes," in *Essays on Actions and Events* (New York: Oxford University Press, 1980).

[2] This section repeats, with minor modifications, material from Hieronymi, "Reasons for Action."

[3] Jonathan Dancy forcefully draws attention to these problems. See Dancy. I here focus on mistakes of fact. One might instead mistake what the facts count in favor of doing. Such cases generate further complication, but, I believe, can be handled in the same way I will propose handling mistakes of fact.

[4] Dancy says, 'there was no reason to do what [she] did, even though [she] did it for a reason' Ibid., 3.

[5] The labels in this area are fraught. Scanlon's use of "operative" differs from that in Joseph Raz,

But, once make that distinction and introduce those labels, we encounter a second, more serious worry for the simple view: Operative reasons cannot *themselves* explain the action, at least in cases of error. We cannot explain Jae's departure by appeal to the fact that the store was closing – because there is no such fact. Something that is not the case cannot explain something that is.[6] We need a fact to explain Jae's departure.

The fact that seems obvious to employ, for this purpose, is the (psychological) fact that Jae *thought* the store was closing. We might, then, abandon the simple view and instead appeal to psychological facts to explain action. In fact, we might appeal to mental states that contain, as their content, the agent's reason for acting.[7] This was Davidson's strategy.

Notice, though, how sharply we thereby separate the reasons that explain Jae's action and Jae's own reasons for acting. The reasons that explain her action are facts about her psychology, while her own reason for acting had nothing to do with her psychology. She did not take facts about her thoughts to count in favor of leaving (as she might if, say, all who did not share the beliefs of the congregation were asked to leave). Rather, she took the imminent closure to count in favor of leaving. And, even though she was mistaken about the closing, she was right to take the closing, rather than her thoughts about it, to be what counted in favor of leaving. Only so can we say that, since the store was not closing, nothing that counted in favor of leaving. And only so can we say that, if the store was closing, there was reason to leave, even if Jae did not know it. If we insisted that our beliefs, themselves, are what count in favor of acting, we would have to say that we do not, by making our beliefs more accurate, improve our information about what we have reason to do. This is unacceptable.[8]

Reasons that count in favor of acting are typically facts about the world at large rather than facts about the actor's own psychology. And yet, in light of our fallibility, it seems that the psychology explains the action. Jae's departure is explained by facts about her psychology, even when she is correct. The imminent closure seems dispensable. And thus it seems, not only that the reasons that explain an action and the

---

Practical Reason and Norms (London: Hutchinson, 1975; reprint, Princeton University Press, 1990), 33.

[6] The fact that p is false can explain q, but *the fact* that p is false is, itself, a truth.

[7] Michael Smith calls these "motivating reasons." Michael Smith, *The Moral Problem* (Oxford: Blackwell, 1994). What Parfit, Dancy, and Schroeder call "motivating reasons," Smith often calls "my normative reason." (Operative reasons face the further requirement that they play a role in explaining action.) I much prefer Smith's use of "motivating."

[8] Even Bernard Williams does not insist that your beliefs are themselves either what counts in favor of action or what you take to count in favor of acting. You have a reason not to drink the petrol, and no reason to do so, even when you believe that it is gin and desire to drink a gin and tonic. (Williams insists you have a reason only if it *possible*, given certain idealizations, for you to believe that you have that reason. See Williams, "Internal and External Reasons.") Thinking that beliefs themselves are what counts in favor of acting is extreme.

agent's reasons for acting are different kinds of things, but also that the second, the agent's reasons, are somehow inert in the explanation. Call this "Dancy's Objection."[9]

Davidson himself eventually raised a second kind of objection for his own view: psychological states that contain considerations that count in favor of acting can explain a person's action, even in cases in which the agent did not act for those reasons. Davidson's example involves a climber who desires to be safe and believes that dropping the rope that is holding his partner would make him safe, and these together so unnerve him that he inadvertently drops the rope. The possibility of such "deviant causal chains" shows that Davidson has not yet answered his question: he has not yet identified the relation between the reason and the action when the reason explains by action by being the agent's reason. The considerations are the agent's reason only if they explain the action "in the right way," as Davidson put it, "through a course of practical reasoning, as we might try saying." He therefore despaired of providing a causal account.[10]

And so our second, Davidsonian, strategy has not succeeded: we have yet to understand the role of the agent's own reason in the explanation of action. Moreover, as noted by Thomas Nagel, that role must, in some way, relate the reasons that explain the action to those (if any) that (in fact) count in favor of acting – lest it turn out that "we don't really act for reasons at all ... we are caused to act by desires and beliefs, and the terminology of reason can be used only in a diminished sense to express this kind of explanation."[11]

A third strategy (one which, I believe, was the target of Davidson's article) would deny that actions are explained in anything like the way we explain other (mere) happenings. According to this third account, the question "Why did Jae leave?" and the question "Why did the computer crash?" bear only surface similarity. If you ask "Why did the computer crash?" you are pursuing an ordinary explanation, asking, in a quasi-scientific way, "How did it come about that the computer crashed?" But, one might think, when we explain – when we make intelligible – a human action, we are engaged in a different sort of project, answering a very different sort of question. We are not asking, "How did it come about that Jae left?" in a quasi-scientific spirit. We are instead seeking to make her action intelligible by asking, "From Jae's point of view, why leave?" That is, to explain action, *qua* action, is not to say how an ordinary event came about, but rather to say what, from the agent's point of view, counted in favor of so acting. Thus the reasons we appeal to, in explaining the action, are the reasons the

[9] See Dancy. I hope the parallel to certain forms of skepticism is clear: if we explain the non-veridical case by appeal to appearances, it seems we no longer have need of reality. Dancy makes the connection in "Arguments from Illusion," *The Philosophical Quarterly* 45, no. 181 (1995): 246–8. Although I see Dancy's worry, I am not, myself, gripped by it. Below I will explain why.

[10] Donald Davidson, "Intending," in *Essays on Actions and Events* (Oxford: Oxford University Press, 1980), 79. Though many have taken the problem of deviant causal chains to set a research agenda (locate the right causal chain), Davidson's anomalous monism bars this route for him.

[11] Thomas Nagel, *The View from Nowhere* (Oxford: Oxford University Press, 1986), 142.

agent might use, in deciding whether to act. It will, of course, be entirely unremarkable that such "explanations," framed as they are from another's point of view, sometimes refer to falsehoods. When they do, then, to avoid confusion, we will mark that fact by saying, for example, "Jae left because *she thought* the store was closing." But, in this context, appealing to falsehoods is not a problem – we are not explaining how something came about, but rather how things appear from a certain vantage. And so the addition of "she thought" does not contribute to the explanation – it is not an appeal to a piece of psychology. It simply makes explicit what is true in any such explanation: it is given from the agent's point of view.[12]

While I have great deal of sympathy for this kind of view, it was a position of this sort that Davidson's article displaced.[13] Davidson, in effect, pointed out that there may be a great many possible answers to the question, "From her point of view, why leave?" which, as things in fact happened, played no role in her leaving – because, for example, she did not notice them. In answering the question, "From her point of view, why do thus-and-such?", we will make intelligible why someone *could* or *would* or *might* so act. We reveal relations of justification that hold between features of the situation. But we have not, thereby, yet done anything to explain what in fact happened. Davidson, in effect, simply reasserted the demand for a more ordinary explanation. The demand seems to me appropriate – and our questions remain outstanding: How do we relate the agent's own reasons for acting to either the reasons that explain her action or the reasons (if any) that in fact count in favor of acting? This is the second difficulty in our catalog: the explanation of action done for reasons.

[12] Dancy seems to adopt this kind of view: "We explain the action by showing that the answer to the ... question [Had things been the way he supposed them to be, would his action have been the one there was most reason to do?] is yes ... . to explain an action is to justify it only in a certain sense" Dancy, *Practical Reality*, 9. Later he says, "The explanation of an action succeeds to the extent that it enables us to see how the agent might have taken certain features of the action as good reasons to do it" Ibid., 95. (Note the "might have." That is the hook for Davidson's objection.) It is worth noting that this view can allow other ways of explaining the event – neural explanations, for example. See Ibid., 176–7. See also "Two Ways of Explaining Action," *Royal Institute of Philosophy Supplements* 55 (2004).

[13] Davidson's explicit target was A. I. Melden, *Free Action* (London: Routledge, 1961). Another was G. E. M. Anscombe, *Intention* (Oxford: Blackwell Publishing Co., 1957). For contemporary versions, in addition to Dancy, see Frederick Stoutland, "The Real Reasons," in *Human Action, Deliberation, and Causation*, eds. Jan Bransen and Stefaan E. Cuypers (Dordrecht: Kluwer Academic Publishers, 1998); Alan Millar, *Understanding People* (Oxford: Oxford University Press, 2004).

The third difficulty

Moving to a third: the standard accounts of what it is to be a reason leave open a problem that is called, by some of us, "the wrong kind of reason problem."[1] Recall that the standard accounts understand "normative" reasons as considerations that count in favor of (or justify, show valuable or correct, or stand in a "normative" relation to) actions or attitudes. But certain considerations seem to *count in favor of* believing or admiring or intending (that is, they bear the same relation, whatever it is, to believing or admiring or intending that reasons for action bear to acting), and yet they seem to be the wrong kind of reasons for the attitude. For example, the fact that it would let you sleep is a reason for believing everything will work out. It surely counts in favor of believing, in just the same way it counts in favor of wearing earplugs or counting sheep. It bears the same relation to believing that it bears to those other activities. But it is the wrong kind of reason for believing. We seem to encounter the same problem for a host of attitudes.[2]

Some attempt to address this problem by identifying, as the right kind of reason, the reasons that would show the attitude to be good, or correct, or fitting as an attitude of that sort. Thus the right kind of reasons for a belief are those that show it to be good as a belief, the right kind of reason for admiration are those that show admiration fitting, the right kind of reason for intending are those that show intending correct, and so on. Such an account must, of course, specify what it is to be "good as" or "fitting" or "correct," if it is to identify the right kind of reason. But there are two further, less obvious, challenges such an account must face.

First, such accounts will identify reasons of the right kind with *good* reasons, but the distinction between reasons of the right and wrong kind seems orthogonal to the distinction between good and bad. While the fact that it would help me sleep is the wrong kind of reason for believing everything will work out, the fact that I am a Capricorn and my stars are aligned is just a *bad* reason – it is not a reason of the wrong kind. We need a way to understand bad reasons of the right kind.

[1] I am understanding the wrong-kind-of-reasons problem in the way I have elsewhere understood it, and I will use "wrong kind" accordingly. For a short summary, see Pamela Hieronymi, "The Use of Reasons in Thought (and the Use of Earmarks in Arguments)," *Ethics* 124, no. 1 (2013).

[2] The problem seems to be that the relation in which a reason stands when it counts in favor of an action (showing the action good or worth doing) is not the relation in which a reason stands when it counts in favor of an attitude (showing something about the target or content of the attitude). In the former case, "counts in favor of" means something like "shows something good about bringing about," but in the latter, "counts in favor of" just means "is a reason for."

One might respond by claiming that reasons of the right kind are those the person *took* to show the attitude as good of its kind. Thus, the fact that the stars have aligned is a reason of the right kind so long as the thinker takes it to show that the belief is good of its kind, but it is a bad reason of the right kind of the thinker is mistaken.

This response addresses the challenge by attributing to the thinker thoughts about what makes beliefs good, *qua* beliefs.[3] We would be unable to draw the distinction for any thinker who lacks the concept of belief.

Even if we accepted this cost, we face another difficulty: While it is criticizable, and sometimes even irrational, to believe for bad reasons, it seems (at least to many of us) *impossible* to believe for reasons of the wrong kind. That is, it seems, at least to many people, that you cannot believe at will.[4] It similarly seems impossible to admire or resent for reasons of the wrong kind. An account that identifies reasons of the wrong kind as reasons that fail to show something good of its kind will leave this unexplained, because failing to show a thing good of its kind is not generally a bar to employing a reason. To illustrate: I can make a move in our chess game not because it would be good *qua* chess move but because it will end the game so we can all finally leave. Even though I do not think this reason shows the move good *qua* chess move, I have no difficulty acting on it. In contrast, even if I think the importance of sleep, on this occasion, massively outweighs the good of maintaining a proper epistemic state, I cannot believe in order to get a good night's sleep. And, just as importantly, even if I mistakenly think that the fact that it would help me sleep shows the belief good *qua* belief (because, perhaps, a good night's sleep will help with tomorrow's scientific investigations), I still will not be able to believe for this reason.

[3] Alternatively, one could posit a mechanism that does this work. See Nishi Shah and J. David Velleman, "Doxastic Deliberation," *Philosophical Review* 114, no. 4 (2005). I reply at Pamela Hieronymi, "Controlling Attitudes," *Pacific Philosophical Quarterly* 87, no. 1 (2006): footnote 4, by essentially making the point in the next paragraph in the main text.

Another strategy, pursued by Kurt Sylvan, would identify the right kind of reason as those competently *treated* as reasons, where competence is understood dispositionally. This successfully avoids the over-intellectualization problem and would allow for some fallibility and occasional performance failures. However, because Sylvan understands "reasons" as good reasons (what he calls "objective reasons"), the view will rule out cases in which a person reasons reliably badly: If I am reliably disposed to treat the fact that I am a Capricorn as a reason to draw conclusions about my fate, I will not be manifesting a competence. Perhaps this strategy could be modified to claim that the right kind of reasons are those that are treated in the way that they would correctly be treated if they did show the attitude fitting or correct. This would avoid the over-intellectualization problem and allow bad reasons (though it would run into trouble with the next point in the main text). The difficulty will be in identify "the way they would correctly be treated, if they did show the attitude fitting." I suspect that "way" is "as bearing on the relevant question." See Kurt Sylvan, "What Apparent Reasons Appear to Be," *Philosophical Studies* 172 (2014).

[4] For an argument that it is impossible, see Hieronymi, "Controlling Attitudes"; "Believing at Will," *Canadian Journal of Philosophy, Supplementary* 35 (2009).

Alternative

With these three difficulties in mind (the mystery of the "normative" relation, the difficulty of identifying the role of the agent's reason in the explanation of actions done for reasons, and accounting for the difference between the right and wrong kind of reason for certain attitudes), I would suggest an alternative account of reasons, one which re-arranges the pieces and thereby rearranges the philosophical tasks. When trying to understand reasons, we should not start with the fact that reasons explain, or justify, or count in favor of, or motivate events, states of affairs, attitudes, or actions. They do all of these in virtue of a further, more fundamental fact about them. I suggest we begin instead with this thought: Reasons are items in pieces of actual or possible reasoning. Reasoning is thought organized in a certain way: directed at a question or conclusion. Thus, I would suggest, reasons are considerations that either bear or are taken to bear on a question.

An important thing to note: Reasoning can be wrong, mistaken, off, and still be reasoning. Thus, on this way of understanding reasons, bad reasons are still reasons – they are just bad reasons. Good reasons are considerations that actually bear, or that are correctly taken to bear, on a question. Bad reasons are considerations that are taken to bear on a question but are not good reasons.[1]

Another important thing to note: taking is not believing. This alternative account does not claim that reasons are considerations *believed* to bear on a question. To take a consideration to bear on a question is not to form a belief about the consideration, the question, and the "bearing-on" relation. Rather, to take a consideration to bear on a question is to employ that consideration in addressing the question. Again, reasons are items in pieces of reasoning.

Finally, reasoning is organized thought, not explicit deliberation. Explicit deliberation is a conscious activity that unfolds across time. Organized thought need not be. I can take reasons to bear on, or to settle, a question without explicitly deliberating about that question.

The most important change, in moving to this proposed alternative, is this: Considerations no longer become reasons in virtue of some relation in which they stand to an event, a state of affairs, an action, or an attitude – whether explanatory or "norma-

[1] So, Jae had a reason to leave, but it was not a good reason, because the store was not, in fact, closing. The distinction between good and bad admits all the different layers that the distinction between justified and unjustified admits: correctly taken to bear, given omniscience and omnibenevolence, or given goodwill and what the thinker believes at the time, or given what the thinker ought to have known, had they exercised due care, or ... and so on.

tive." Instead, considerations become reasons in virtue of their relation to a question. With this alternative in view, I hope the idea of relating *considerations* directly to *events* or *states of affairs* – even psychological states and events that are actions – will seem odd, a kind of unholy juxtaposition of the rational and the empirical.[2] But, more to the point, by relating reasons first to questions, we thereby require questions to mediate between considerations, on the one hand, and, on the other, the actions or attitudes they might explain, justify, count in favor of, show correct, or ground. This mediation by questions is the most important change, because it allows us – in fact, it requires us – to bring rational agency into view: it is the rational agent who, by settling questions, by concluding or deciding, forms attitudes and sets themselves to act – sets themselves to bring about events or states of affairs. It is thus the agent, the thinker, who mediates between considerations, on the one hand, and states of affairs or events, on the other. Views that relate considerations directly to attitudes or actions, even by appeal to multi-place relations that include the agent, thereby obscure the agent's role – they obscure the activity of the thinker in concluding or deciding or committing. The most important contribution of this alternative account is to bring rational agency (reasoning, concluding, deciding) into view. I hope now to show how doing so helps to address the difficulties we have considered.

[2] One might recall the quote from Strawson about the difference between explanatory and causal relations.

Identifying the wrong kind of reasons and the "voluntary"

Let us start with the wrong kind of reasons problem. Notice, first, that certain states of mind (e.g., belief, intention, admiration, resentment) can relate to questions in two distinct ways. A state of mind sometimes appears in the *content* of a question, as part of what the question is about. We can ask why she believes her country is less safe, or when he became so angry, or why they admire him so much. But certain states of mind relate to questions in a different, more direct – or, perhaps, more indirect – way. Consider the relation between the question of whether the butler did it and the belief that the butler did it. By settling the question, you form the belief. But the question is not about your belief. It is about things at some distance from you: the butler and his crime. Still, by settling it positively, you make something true of yourself – right here at home, so to speak. You make it the case that you believe the butler did it. The relation between the question and the state of mind seems indirect if you consider the question's content: the question is not about the state of mind. But it seems direct if you consider agency: by settling the question positively, one *therein* believes.

I would suggest that we understand certain states of mind – most centrally, belief and intention – as themselves forms of question-settling. It is this form, I think, that gives applicability to the request for one's reasons.[1] But in saying this, in saying that to believe P, for example, is to settle the question of whether P, or that to intend to x is to settle the question of whether to x, I do not mean to posit a new, independent psychological event or activity, the settling of a question, that somehow accompanies believing or intending. Rather, I mean to claim that belief, intention, and the rest, are, themselves, helpfully thought of as question-settlings; question-settling is something like a genus into which these attitudes fall as species.

If we see these attitudes as forms of question-settlings, then we can both distinguish the right from the wrong kind of reasons for them and say why they are not voluntary – in fact, we uncover a useful characterization of what "voluntary" means, in this context.

To start, we can distinguish the right from the wrong kind of reason. The right kind of reasons for an attitude that is a question-settling are considerations that bear or are taken to bear on the relevant questions. Reasons of the wrong kind manage to count

[1] This claim will be more fully defended in a manuscript currently in progress. See also "Two Kinds of Agency," in *Mental Actions*, eds. Lucy O'Brien and Matthew Soteriou (Oxford: Oxford University Press, 2009).

in favor of the attitude in some other way – typically by showing the attitude in some other way good or useful or worth having.[2]

We can also give a useful characterization of what we might mean by "voluntary," and we can see why believing is not voluntary in this sense.[3] An activity is voluntary, in the relevant sense, if it can be done for *any* reason that you take to show it worth doing. You can raise your right hand, run for office, or plant azaleas for any reason that you think shows it worth doing – to win a bet, or make a joke, or make a point. In contrast, you cannot believe something (for example, that the butler did it) in order to win a bet, make a joke, or make a point – even if you think it would be worth doing. You can only believe what you take to be true. We can thereby specify the sense in which ordinary actions are voluntary while believing is not.

Finally, we can say *why* believing is not voluntary, in this sense: You might find yourself with reasons that show believing P good to do that you do not take to bear on whether P. You could get a good night's sleep if you could believe everything will work out, but you do not take the possibility of a good night's sleep to show that everything will work out – you take it to show, instead, that it would be good to believe that. But the question of whether everything will work out and the question of whether it would be good to believe everything will work out are different questions, and you cannot settle a question for reasons you do not take to bear on it.[4]

Why can you not settle a question for reasons you do not take to bear on it? Because, if you settle a question for a reason, you have *therein* taken the reason to bear on the question. And so, as a conceptual matter, you cannot settle a question for a reason that you do not take to bear on it. Thus, if you find yourself with reasons that you take to show believing P worth doing (or a belief that P worth having) that you do not take to bear on whether P, you will find yourself with reasons that you take to show believing worth doing, but you will not be able to believe for those reasons. And so it is that belief is non-voluntary: you cannot believe for any reason you take to show believing worth doing.

Perhaps surprisingly, just the same is true of intention. You might have reason that you take to be sufficient reason to *intend* to x– reason enough to house the intention – but that you do not take to be reason enough to x– not reason enough to act. Perhaps you have no intention of marrying your partner, and they are unhappy about this fact. Because you like to please your partner, you would be happy to house the intention – so long as you do not need to go through with the marriage. You are out of luck. In order to intend to marry, you have to decide to marry – to intend, you must settle the question of whether to *act*, not just the question of whether to intend. And so, even though you take yourself to have reason enough to intend, you will not be able to intend. Somewhat surprisingly, then, although you can *act* at will – though you can act

[2] See Hieronymi, "The Wrong Kind of Reason."

[3] This section, about this specific sense of "voluntary," repeats, with minor modification, some material found in "I'll Bet You Think This Blame Is About You," (2019).

[4] This argument is made, at length, in Hieronymi, "Controlling Attitudes" and "Believing at Will."

for any reason you take to show the action sufficiently worth doing – you can no more intend at will than you can believe at will.[5] While actions are voluntary, intentions are not.

The same is also true of a wide range of attitudes – of any attitude that manifests our take on the world, on what is true, important, worthwhile, insulting, wonderful, horrifying, trustworthy, impressive, and so on, for which we can be asked our reasons. Such attitudes must be non-voluntary, in the sense just explained, in order to play the roles they play and bear the significance they bear in our lives. If a state of mind is voluntary, you can do it any reason you take to show it worth doing – you can, for example, imagine a red circle for any reason you take to show it worth doing. But if a state of mind is voluntary in this way, it will not reveal your take on what is true, or important, worthwhile, insulting, and so on. Instead, like an ordinary action, it reveals your take on what is worth doing – in particular, it reveals your take on whether imagining a red circle is worth doing.

We have just connected questions of voluntariness with the wrong-kind-of-reason problem: Attitudes that are non-voluntary, in sense in which believing is non-voluntary, are also, and *therefore*, subject to a wrong-kind-of-reason problem: you might find yourself with reasons that you take to show them worth having that you do not take to bear on the relevant questions.

[5] It is difficult to generate the problem for intention, because there are very few constraints on the reasons for which one can act (most any consideration could, in principle, bear on the question of whether to x), and it is possible to act as a way of making yourself intend. (If you are unhappy that I have no intention to attend your party, I can decide to attend your party in order to keep you happy – even if what you really care about is my intention, not my attendance.) In fact, the case of pleasing your partner is not the exactly a case of the wrong kind of reason for intending, because, if you thought that housing the intention were reason enough to *marry*, you could decide to marry in order to have the intention. The reason *bears* on the question, but you do not take it to be *sufficient* reason to settle the question. In contrast, the Toxin Puzzle case (Gregory Kavka, "The Toxin Puzzle," *Analysis* 43, 1983) and the original case of Mutual Assured Destruction are ones in which the reason to intend does not bear on the question of whether to act, because the reason to intend disappears before the time of action, and this is known in advance. These cases present reasons that are genuinely of the "wrong kind".

The full argument that you cannot intend at will appears in Hieronymi, "Controlling Attitudes." The marriage example appears in "Responsibility for Believing," *Synthese* 161, no. 3 (2008); "Reflection and Responsibility," Philosophy and Public Affairs 42, no. 1 (2014); "Forgiveness, Blame, Reasons ... ," in *3am: Magazine*, ed. Richard Marshall (2013).

Explaining actions done for reasons

In addition to clarity about the wrong-kind-of-reason problem, about voluntariness, and about our agency with respect to our attitudes, we also gain some degree of clarity about the role of the agent's reason in the explanation of actions done for reasons. We can adopt an extremely simple, formal account that will explain an action by appeal to what are, from the explainer's point of view, facts, while also both preserving the proper role of the agent's own reasons for acting (if the agent had reasons)[1] and relating the agent's reasons to the reasons (if any) that in fact count in favor of acting.

The account is embarrassingly simple: We explain events that are actions done for reasons by appeal to the following complex fact: the agent took certain considerations to settle the question of whether to act, therein intended so to act, and successfully executed that intention in action.

Using this form, we answer the ordinary explanatory question, "How did it come about that Jae left the store?" by appealing, in part, to the fact that Jae settled a *different* question – the question of whether to leave. To answer *our* explanatory question, we appeal to the fact that Jae settled *her* practical question for her (operative) reason. Her operative reason thus appears in our explanation, but it appears *as* her operative reason, bearing, for her, on her question. Following Davidson's intuitions, we have explained the action by providing ourselves with something like "a course of practical reasoning" (albeit a very short one).

We have also avoided Davidson's criticisms: We have done more than make the action intelligible from Jae's point of view. We have claimed that certain considerations were those for which Jae, in fact, formed an intention, which intention she executed in the event that was the action. We have, I think, satisfied the demand for a more ordinary form of explanation.

Relatedly, the account avoids the possibility of deviant causal chains: the agent, for certain reasons, settles the question of whether to act, *therein* intends to act, and executes *that* intention in the event that is the action. The connections are too tight for deviance.[2]

Moreover, the account also provides a fairly clear view of the relation between the reasons (if any) that in fact counted in favor of leaving and the reason that explains

[1] The account accommodates action for no (particular) reason (by allowing that we can settle a question for no reason).

[2] One might object: That fact that it does not allow for deviance shows that the account is not explanatory. It merely provides an *analysis* of action done for reasons. No explanation will be given until the pieces of this analysis are filled in. This is an interesting objection. For now, I will simply note that

the action: The complicated fact that explains the action includes within it the fact that the agent *treated* certain considerations as reasons "in the standard normative sense." It thereby addresses Nagel's concern.

One might still harbor Dancy's worry: the agent's own reason – the imminent closure – seems dispensable. But once we have shown the role it plays, I think we need not be troubled. It seems appropriate that the agent, or the agent's activities, should, so to speak, "stand in" for those (purported) facts that the agent takes to be reason-giving. It is the agent, not the facts that call for action, that brings the action to be.[3]

We should notice, though, that not all of the reasons that might explain an action fit into this form. In fact, not all the reasons that both explain and *justify* actions will fit. This is as it should be. The question of whether to act and the question of why someone acted as they did or whether they acted well or as they ought are different questions, and we should expect that we can sometimes answer the latter without answering the former. For example, the fact that he was deceived, or the fact that she was innocently unaware, might both explain and justify an action done for reasons, but, of course, neither of these were the agent's reason for acting. And, we sometimes explain an action (even our own, current action) by setting it in a context that makes it intelligible, without providing the agent's reasons. If asked why I am breaking eggs, I might explain that I am in the middle of making an omelet. I have made myself intelligible to you. But I doubt I have provided you with my reason for breaking the eggs: my own reason for acting cannot, I think, be the fact that I am already in the process of so acting. (We might also wonder whether explanations such as "she was hungry" or "I just felt like it" are functioning to give the agent's own reason or are rather simply placing in context. It may be an open question or perhaps even indeterminate in certain cases.)

By more clearly separating the practical question of whether to act from the justificatory question of whether someone acted as they ought or had reason to, we easily allow for the many different layers of which justification admits: We can ask, did the person do as they ought or had reason to, *given what they knew at the time*? Or, given the facts *they did not know but ought to have known*? Or, given things *as they in fact were*? Each can receive a different answer. In fact, we leave open the possibility of justifying or showing correct (or beautiful) things other than actions and attitudes done for reasons: as, it seems, we should.

I have provided, for action done for reasons, something like the following explanation of how it came about that the Sox won: by the end of the game, they had scored more runs than their opponent. I agree this is an uninformative explanation, perhaps no explanation at all. But, if we hope to explain the win (rather than, say, movements of humans on a field), any (further) explanation must fill out this form. Likewise, I will be satisfied if it is agreed that, if we are to explain action in a way that preserves the role of the agent's reason for acting, the explanation should fit into the form or analysis here proposed. Thanks to Simon Rippon for this objection.

[3] Much of the above sub-section again repeats, with minor modification, material found in Hieronymi, "Reasons for Action."

The mysterious "normative" relation

Finally, let us turn to the category of the so-called "normative." The word came into such prominent use in the philosophical literature, I believe, after the publication of Christine Korsgaard's *Sources of Normativity*.[1] In that work, Korsgaard invites her reader to choose for themselves "the normative word" – the word that indicates, to the reader, that it would be incoherent to continue practical deliberation, incoherent to keep asking whether you really *must*, for example, tell the truth (that is, whether you really should, or ought to, or have most reason to, tell the truth, or whether telling the truth would be the best thing, etc.). For Korsgaard's argumentative purposes, she explicitly wants the word to slip between these different ideas. But slipperiness now seems its legacy, as a piece of philosophical jargon. Depending on the writer, "normative" may now mean "having something to do with reasons," or with values, or with standards, or with questions of appropriateness, or even with blame or the "reactive attitudes." Worse, one can now find the word qualifying any of these – so that, in addition to reading about "normative reasons," one can now read about "normative *standards*," or even "normative criticism." I have heard the phrase "non-normative good." In fact, given that the distinction between explanatory and "normative" reasons need not track the distinction between good and bad reasons, someone might like to refer to the "*normative* normative reasons." The situation has become absurd.

My own preference is to simply avoid the word, when speaking in my own voice, and to insist on more precision. In doing so, we may lose touch with what some people think of as a pressing philosophical project – an outcome I would welcome. To explain:

As noted at the start, those considering reasons as such tend to see them as considerations standing in some relation (perhaps a multi-place relation), and tend to divide them, broadly, between explanatory and "normative." Many then seem content to treat the relation in which explanatory reasons stand to that which they explain (the "explanatory relation") as primitive but find "normative" relations somewhat mysterious, themselves in need of an explanation. That is to say, they think we need to explain why or how certain things make other things good or bad, correct or incorrect, important or unimportant, apt or inapt, obligatory or permissible.

Or, better, they think we need to explain *what it is* for certain things to make other things good or bad, correct or incorrect, important or unimportant, required

[1] Christine M. Korsgaard, *The Sources of Normativity* (Cambridge: Cambridge University Press,

or permissible, and so on. A satisfying explanation of "normativity" must do more than restate the case for specific answers to specific questions – do more than say why, for example, it is important to brush your teeth or why you are obliged to keep your promises or why it is inapt to end the song on that chord. To answer specific questions is, after all, simply to give further considerations that count in favor of brushing your teeth or keeping your promises or resolving the chord, not to illuminate the "normativity" of those considerations. Nor, it might seem, will it do simply to class some of these answers into domains and notice their structure or similarity – to say that moral obligations arise in this way, while prudential requirements arise in that – because, again, such a grouping will do nothing to explain the "normativity" of the domain identified.[2] And so it might seem that we must answer a higher-order question about "normative" relations, in general – what they are, how they hold, and why we are entitled to reason with them. Such an explanation is very difficult to give. However, absent one, it can seem we are forced to choose between either simply granting a strange new primitive or else either discrediting or reinterpreting the thoughts that traffic in these terms.

If we instead adopt the view here suggested, the philosophical tasks rearrange themselves. We will no longer think that reasons stand in either "explanatory" or "normative" relations to events or states of affairs. They stand, rather, in relation to questions. The relation in which a reasons stands to a question is neither explanatory (and so, somehow, unproblematic) nor "normative" (and so, somehow, problematic). It is, rather, the *question* that is explanatory or otherwise.

In fact, it is now difficult to see how to draw the distinction between "explanatory" and "normative." The question "Why did the engine fail?" seems explanatory, and the reasons that bear on it might be called explanatory reasons.[3] But is the question of whether the butler did it an explanatory or a normative question? Reasons for or against believing the butler did it – what some would call "normative reasons" for or against this belief – bear on the question, "Did the butler do it?", but their relation to *that* question seems no more (nor less) "normative" than that which holds between the question "Why did the engine fail?" and the considerations that bear on it. If the considerations that bear on whether to take an aspirin or resolve the chord are "normative" in some further sense, that might be in virtue of the fact that, in asking those questions, I am asking what to bring about rather than what is the case. But now we have drawn a distinction between (what is sometimes called) the "practical" and

1996).

[2] Scanlon appeals to domains in just this way, and he accepts a primitive. See Scanlon, *Being Realistic About Reasons*.

[3] There is complication here. The extreme heat explains the failure, but it seems to be the *answer* to the question, rather than a consideration that bears on it. So perhaps the reasons that bear on explanatory questions are not explanatory reasons – distancing us even further from the standard account. (I owe a debt to someone for this objection, and a further debt to them for having lost track of who it was.)

the "theoretical" or "epistemic" – while, on many uses, epistemic reasons are normative reasons.

We might, of course, stipulate some class of questions as the "normative" ones, but I do not see which are the obvious candidates – nor, more importantly, do I see why we would want to do so.

To be sure, we will not avoid the philosophical task of understanding what it is for something to be good or bad, correct or incorrect, justified or unjustified, obligatory or permissible. But once we give up the idea that the "explanatory" relation is unproblematic while "normative" relations require explanation, we may not feel the same need to give an entirely general account. The appeal to domain-specific answers – to different answers for music, politics, medicine, epistemology, and metaphysics – may no longer seem so dissatisfying (even if we will want, reasonably, to consider their interaction).[4]

Returning now, briefly, to the no-background-conditions view: Once we appeal to questions the view seems natural, because, typically, by showing that or how some other consideration bears on a question, a consideration will itself, thereby, bear on that question. Suppose you are wondering why the fact that she is exhausted is a reason to help her, and you are told that it is because you would want help, if you were exhausted. You accept this explanation. Now, it seems, you will think the fact that you would want help, if you were exhausted, bears on the question of whether to help her, given that she is exhausted. If you think of reasons as considerations that bear on questions, it will seem that, typically, a reason that explains why another consideration is a reason to act will, itself, become part of the reasons for action – because it thereby bears on the question of whether so to act. (There are, however, interesting exceptions, such as in games or institutional roles or reductio arguments.)

[4] One might look for an explanation or elaboration of the "bearing on" relation. Again, I doubt we will find more than domain-specific answers. The fact that the butler had ready access to the home bears on the question of his guilt. If we want to explain why that fact bears on that question, we will point to facts about the crime and what was required to commit it. In giving that explanation, we will have taken for granted other "bearing on" relations, which may, in turn, be explained. I do not see a problem here, nor a general (rather than domain-specific) question that remains mysterious. (I am thus sympathetic to the views found in Ibid. and Sarah Buss, "Against the Quest for the Source of Normativity" [In progress].)

Conclusion

The case for starting with the use of reasons in thought – for thinking of reasons as items in pieces of actual or possible reasoning – is large but cumulative. By doing so, we can avoid the wrong-kind-of-reason problem, understand why beliefs and other attitudes are not voluntary, and recast certain metaethical worries. Elsewhere I have suggested that we also provide ourselves a with way to understand our answerability for our actions and attitudes and a way to model of a central form of weakness of will.[1] However, the greatest benefit lies in avoiding the difficulty that underlies the rest: By modeling reasons as considerations standing in relation to events or states of affairs, the more standard accounts obscure the role and activity of the thinker. It remains mysterious what we do with reasons, how anyone acts on a reason, or how anything is anyone's reason for believing, resenting, trusting, or acting. Yet rational agency, the activity of settling a question, is, itself, something we would like to understand, explain, and evaluate. We help ourselves by exposing it.

What we will not do, I think, by thinking about reasons as such, is to discover which are the *good* reasons. If we think about reasoning, we might learn something general about reasons and agency. But even if we were to understand what makes for good reasoning, I am doubtful that understanding good reasoning, as such, will help us understand very much about how to live or how to treat other people or which actions are good – any more than it will help us to understand very much about how to bake a cake or which things are beautiful. But, at this point, that is mere conjecture.[2]

[1] See Pamela Hieronymi, "The Will as Reason," *Philosophical Perspectives* 23 (2009); "Reflection and Responsibility."

[2] As noted, this chapter draws heavily on earlier work, and gratitude shown there should be repeated here. In addition, thanks are due to John F. Horty, audiences at Rice University, the University of Maryland, College Park; the *Ethics of Belief* conference at Harvard University; and members of the Ethics Workshop at UCLA. Finally, Kurt Sylvan provided extremely helpful comments in his role as editor.

References

Anscombe, G. E. M. *Intention*. Oxford: Blackwell Publishing Co., 1957.

Broome, John. "Reasons." In *Reason and Value: Themes from the Moral Philosophy of Joseph Raz*, edited by R. Jay Wallace, Phillip Pettit, Samuel Scheffler, and Michael Smith, 28–55. Oxford: Clarendon Press, 2004.

Buss, Sarah. "Against the Quest for the Source of Normativity." (In progress).

Dancy, Jonathan. "Arguments from Illusion." *The Philosophical Quarterly* 45, no. 181 (1995): 421–38.

———. *Practical Reality*. Oxford: Oxford University Press, 2000.

———. "Two Ways of Explaining Action." *Royal Institute of Philosophy Supplements* 55 (September 2004): 25–42.

Davidson, Donald. "Actions, Reasons, and Causes." In *Essays on Actions and Events*, 3–19. New York: Oxford University Press, 1980.

———. "Intending." In *Essays on Actions and Events*, 83–102. Oxford: Oxford University Press, 1980.

Finlay, Stephen. *Confusion of Tongues: A Theory of Normative Language*. New York: Oxford University Press, 2014.

Fogal, Daniel. "Reason, Reasons, and Context." Chap. 4 In *Weighing Reasons*, edited by Errol Lord and Barry Maguire. New York: Oxford University Press, 2016.

Hieronymi, Pamela. "Internal Reasons and the Integrity of Blame." [Unpublished Manuscript, https://ucla.box.com/v/HieronymiIntegrityOfBlame].

———. "The Wrong Kind of Reason." *The Journal of Philosophy* 102, no. 9 (September 2005): 1–21.

———. "Controlling Attitudes." *Pacific Philosophical Quarterly* 87, no. 1 (March 2006): 45–74.

———. "Responsibility for Believing." *Synthese* 161, no. 3 (April 2008): 357–73.

———. *"Believing at Will."* Canadian Journal of Philosophy, Supplementary Volume *35 (2009): 149–87.*

———. "Two Kinds of Agency." In *Mental Actions*, edited by Lucy O'Brien and Matthew Sorteriou, 138–62. Oxford: Oxford University Press, 2009.

———. "The Will as Reason." *Philosophical Perspectives* 23 (2009): 201–20.

———. "Reasons for Action." *Proceedings of the Aristotelian Society* 111 (2011): 407–27.

———. "Forgiveness, Blame, Reasons ..." In *3am: Magazine*, edited by Richard Marshall, 2013.
———. "The Use of Reasons in Thought (and the Use of Earmarks in Arguments)." *Ethics* 124, no. 1 (October 2013): 114–27.
———. "Reflection and Responsibility." *Philosophy and Public Affairs* 42, no. 1 (2014): 3–41.
———. "I'll Bet You Think This Blame Is About You." In *Oxford Studies in Agency and Responsibility, Volume 5: Essays on Themes from the Work of Gary Watson*, edited by Justin Coates and Neal Tognazzini, 60–87. New York: Oxford University Press, 2019.
Horty, John F. *Reasons as Defaults*. New York: Oxford University Press, 2012.
Kant, Immanuel. *Groundwork of the Metaphysics of Morals*, edited by Mary Gregor. Cambridge: Cambridge University Press, 1997.
Kavka, Gregory. "The Toxin Puzzle." *Analysis* 43 (1983): 33–36.
Korsgaard, Christine M. *The Sources of Normativity*. Cambridge: Cambridge University Press, 1996.
Mackie, J. L. *Ethics: Inventing Right and Wrong*. New York: Penguin Books, 1977.
Melden, A. I. *Free Action*. London: Routledge, 1961.
Millar, Alan. *Understanding People*. Oxford: Oxford University Press, 2004.
Nagel, Thomas. *The View from Nowhere*. Oxford: Oxford University Press, 1986.
Pritchard, H. A. "Does Moral Philosophy Rest on a Mistake?" *Mind*, 21, n. 81 (1912): 21–37.
Raz, Joseph. *Practical Reason and Norms*. London: Hutchinson, 1975. Princeton University Press, 1990.
Scanlon, T. M. *What We Owe to Each Other*. Cambridge: Harvard University Press, 1998.
Scanlon, T. M. *Being Realistic About Reasons*. Oxford: Oxford University Press, 2014.
Schroeder, Mark. *Slaves of the Passions*. Oxford: Oxford University Press, 2008.
Shah, Nishi, and J. David Velleman. "Doxastic Deliberation." *Philosophical Review* 114, no. 4 (October 1, 2005): 497–534.
Skorupski, John. *The Domain of Reasons*. Oxford: Oxford University Press, 2010.
Smith, Michael. *The Moral Problem*. Oxford: Blackwell, 1994.
Stoutland, Frederick. "The Real Reasons." In *Human Action, Deliberation, and Causation*, edited by Jan Bransen and Stefaan E. Cuypers, 43–66. Dordrecht: Kluwer Academic Publishers, 1998.
Strawson, Peter F. *Analysis and Metaphysics*. Oxford: Oxford University Press, 1992.
Sylvan, Kurt. "What Apparent Reasons Appear to Be." *Philosophical Studies* 172 (2014).
Williams, Bernard. "A Critique of Utilitarianism." In *Utilitarianism: For and Against*, edited by J. J. C. Smart and Bernard Williams, 75–150. Cambridge: Cambridge University Press, 1973.

———. “Internal and External Reasons.” In *Moral Luck*, 101–13. Cambridge: Cambridge University Press, 1981.
———. “Persons, Character, and Morality.” In *Moral Luck*, 1–19. Cambridge: Cambridge University Press, 1981.

25. Normative nonnaturalism[4]

[4] Prefatory note from Ruth Chang: Two weeks before his sudden death on January 1, 2017, Derek Parfit contacted me about his contribution to this volume. He said he planned to discuss two topics: his theory of normative practical reasons and his metanormative view about the nature of normative facts and claims. He also said that he expected to quote extensively from his already-published three-volume opus, *On What Matters* (Oxford: Oxford University Press, 2011 (vols. 1 & 2), 2017 (vol. 3)).

Derek Parfit (edited by Ruth Chang)

I have tried to honor Parfit's intentions by carving two chapters from *On What Matters*, one on each of the topics he intended to write about, with a minimum of editorial intervention. The following chapter describes and argues for Parfit's metanormative view, a view he was increasingly excited by toward the end of his life because he thought that it was a truth on which seemingly competing metaethical views could – and he hoped would – eventually converge. A companion chapter, 'Objectivism about Reasons', gives Parfit's main defense of the idea that all normative practical reasons are 'value-based' and not 'desire-based'. Both papers are condensed and slightly modified reprints of material from key chapters in On What Matters.

Peter Momtchiloff of Oxford University Press kindly agreed to allow Parfit's material to be re-purposed for this volume on the condition that the authorship of the paper indicated that the material had been edited by me. Needless to say, the work here is entirely Parfit's. We hope that this condensed version of Parfit's arguments will lead readers to investigate the original text.

* * *

> When we claim that some things matter, we might mean only that these things matter to people. Suffering matters, for example, in the sense that people care about suffering. No one doubts that some things matter in this psychological sense. Some things also matter, I believe, in the different, normative sense that we have reasons to care about these things.
>
> – Derek Parfit, *On What Matters*, vol. 3, xv

At the end of his life, Derek Parfit was much occupied with defending the idea that things matter. Suffering matters in the sense that we have reasons to care about suffering and to conduct our lives so as to minimize it. Suffering would matter even if in fact no one actually cared about it. As Parfit wrote, "[t]his view seem[s] to me worth defending because so many people are falsely taught, even at the best universities, that nothing matters in this sense" (vol. 3, 188).

Parfit thought that for things to matter in this reason-implying sense, a non-naturalist view of normativity has to be correct. He develops one such view, Non-Realist Cognitivism (vol. 3, 4), according to which "there are some true claims which are not made to be true by the way in which they correctly describe, or correspond to, how things are in some part of reality" (vol. 3, 4). Paradigmatic cases of such truths are "logical, mathematical, and modal truths, and some fundamental normative truths" (vol. 3, 4). Thus, although there are irreducibly normative truths that things matter, we shouldn't think they are true because they correspond to some bit of reality. Some truths can be true of necessity and not because they describe something in the world. There is no fact in reality that two plus two equals four; there is only the mathematical necessity that it does. Similarly, there is no fact in reality

that we have a reason to care about suffering; there is only the normative necessity that we should. Parfit's Non-Realist Cognitivism allows us to believe that there are irreducibly normative truths without forcing us to hold that there are some mysterious non-natural bits of reality that those truths describe. We can be cognitivists about normativity without being realists about it.

Parfit was delighted to discover that once he gave up the 'realist' part of cognitivism, there was a convergence of sorts between his view and those of certain sophisticated forms of naturalism and expressivism. All three metaethical views, Parfit noted, can accept that that there are or could be irreducibly normative truths of the sort 'suffering matters'. This, for Parfit, was a huge and welcome discovery. Naturalists, non-naturalists, and expressivists could all agree that things matter in Parfit's purely normative, reason-involving sense.

In fact, Parfit argues for two convergences, one on a general formulation of what reasons we have to care about and to do things – what he calls 'The Triple Theory' – and the other on the nature of claims that we have such reasons – Non-Realist Cognitivism. Approached critically, it might be said that in the former case Parfit succeeds by downplaying key aspects of his opponents ethical theories. In the latter case, it might be argued, Parfit succeeds by ignoring key aspects of the view he supposedly champions, at least as it has been traditionally understood. Non-naturalist cognitivism typically holds that reality includes *sui generis*, fundamental non-natural facts that make irreducibly normative claims true. It is true that suffering matters because it is a fundamental fact of reality that it does, and this fact is what makes the claim true. Parfit's nonnaturalist cognitivism rejects these – some would say essential – parts of the view, at least as it has been traditionally understood. By doing so, he accepts a central – perhaps essential – claim that naturalists like Railton and expressivists like Gibbard make, namely that all there is to reality are natural facts. So one question to ask is whether the fact that naturalists and expressivists can allow that there are irreducibly normative truths but no reality that makes them true unmasks them as closet non-naturalists or whether the fact that Parfit allows that there is nothing in reality that makes normative claims true makes him a closet naturalist. Central to answering this question is understanding what is essential to normative non-naturalism. Is it the claim that there are irreducibly normative truths or is it the claim that such truths are made true by the way reality is?

Parfit often said that his main opponent was Normative Naturalism, the view that normative claims describe natural facts and refer to natural properties in the world. The bulk of this chapter is a summary of Parfit's arguments against that view. Of course, such a view is compatible with the idea that normative concepts are irreducibly normative and that there can be irreducibly normative truths employing such concepts. That, indeed, is the sort of naturalist view that Parfit ultimately finds congenial. Normative naturalists are mistaken, Parfit argues, in thinking that normative claims refer to natural properties in the world. But they are correct when they allow that normative concepts *like ought, right, wrong* may not be reduced to natural concepts. Once

a naturalist allows that some concepts are irreducibly normative – they cannot be defined or their meaning fully expressed in non-normative terms – they can also allow that there are irreducibly normative truths. These truths are true not because of the way the world is. They just are true, and if fundamental, true as a matter of normative necessity.

The argument against Normative Naturalism that follows is somewhat lengthy, but it is a condensed summary of an even longer run of argument Parfit had given in an earlier volume of *On What Matters*. The reader should remember that Parfit is here attempting to put to rest what has been for many decades and may in some circles still be the leading metanormative view, the view that our normative claims refer to natural facts and properties. The chapter ends with some brief remarks about how Parfit's Non-Realist Cognitivism might be thought to converge with sophisticated forms of naturalism and expressivism.

We start with Parfit's map of the metaethical landscape.

* * *

1 Meta-ethics

> By asking certain questions we can roughly distinguish several views that are metaethical in the sense that they are about the meaning and truth of moral claims, and of other normative claims.
>
> (vol. 3, 55)

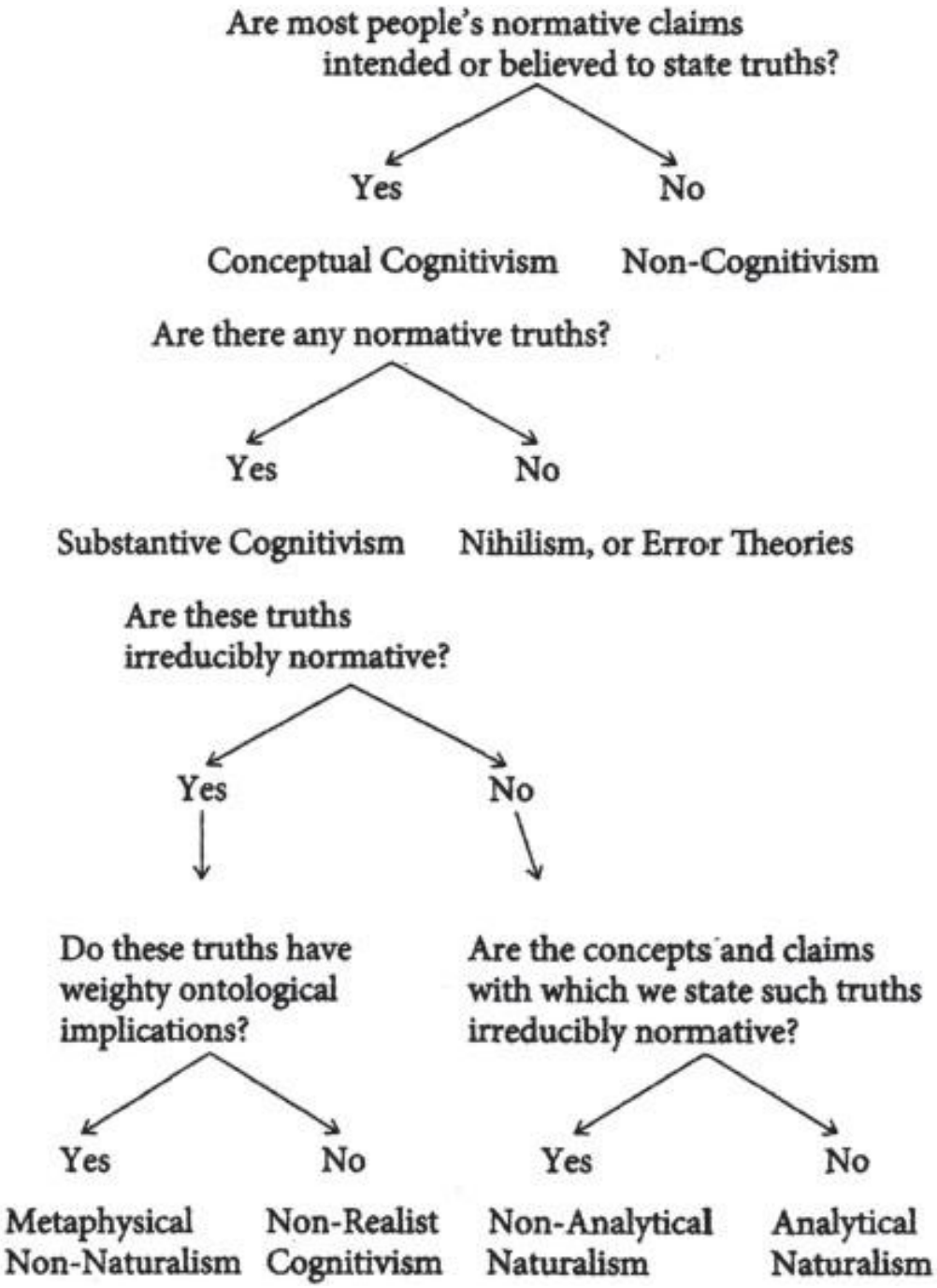
[A chart showing the relations between different metaethical views follows]: (vol. 3, 55)

Since most of us believe that some normative claims are true, NonCognitivists ought to revise their view, by becoming Conceptual Cognitivists. If these people continue to believe that there are no normative truths, they should also become Nihilists, or Error Theorists ...

Of the Cognitivists who believe that there *are* some normative truths, some are Normative Naturalists. Normative truths, these people believe, are like other truths about the natural world which might be empirically discovered, in the sense that some partly observable things or events might give us evidence for or against beliefs in these truths.

According to *Analytical* Naturalists, normative concepts and claims can all be defined or restated in non-normative ways. This view is fairly plausible when applied to some normative concepts and claims. When some people say, for example, that some act would be rational, these people may mean only that this act would achieve the agent's aims. That is not a normative claim. Some evaluative claims could also be plausibly restated in non-normative naturalistic terms. When people say that something tastes good, or that some medicine is the best, these people may mean only that they like this taste, or that this medicine is the safest and most effective. These are not normative claims.

Analytical Naturalism cannot be plausibly applied, however, to some other normative concepts and claims... . [S]ome people use the concept of a purely normative reason. This concept cannot be helpfully defined by using other words. Some fact is a reason, or gives us a reason, we can say, when this fact *counts in favour* of our having some belief or desire, or acting in some way. But that is merely any other way of saying that this fact is a reason, or gives us a reason. Nor can most moral concepts or claims be plausibly defined or restated in non-normative ways. Though there have earlier been some Analytical Moral Naturalists, these people's view are now rightly regarded as too implausible to be worth discussing. No one, for example, now believes that, when


Are most people's normative claims intended or believed to state truths?
Yes
No
Conceptual Cognitivism
Non-Cognitivism
Are there any normative truths?
Yes
No
Substantive Cognitivism
Nihilism, or Error Theories
Are these truths irreducibly normative?
Yes
No
Do these truths have weighty ontological implications?
Are the concepts and claims with which we state such truths irreducibly normative?
Yes
No
Yes
No
Metaphysical Non-Naturalism
Non-Realist Cognitivism
Non-Analytical Naturalism
Analytical Naturalism

pg-368.jpg

most people say that some act would be morally right, they mean that this act would minimize suffering, or would be an act of which most people would approve.

Non-Analytical Naturalists believe that some normative concepts and claims are *irreducibly normative*, in the sense that they cannot be defined or restated in non-normative ways. But these Naturalists also believe that, when such normative claims are true, they state natural facts. On such views, though the phrase 'morally right' does not mean 'would minimize suffering,' or'would be approved by most people,' the fact that some act is right might be the same as the fact that this act minimizes suffering, or is an act of which most people would approve.

According to Non-Naturalists like me, irreducibly normative claims could state only irreducibly normative truths. We should distinguish, however, between at least two kinds of normative truths. Some truths are normative in *norm-implying* or *rule-implying* senses. My examples are the facts that, in some community, certain acts are illegal, or would contravene some widely accepted moral norm, or some professional code, or code of honour, or the rules of etiquette, or linguistic rules about what would be incorrect spellings or misuses of some word. These normative truths are, I believe, natural facts in the sense that they can be explained in naturalistic terms. We can describe how people can create some legal system, and how some law can then be passed which prohibits certain acts. We can then say: 'That's *what it is* for these acts to be illegal.' We can make similar claims about how people can make it true that, in some community, there are some other widely accepted norms or rules about etiquette, or about how our words should be spelt. These normative truths are also natural facts in the closely related sense that they might be empirically discovered. We could often find out whether, in some community, certain acts are illegal, or would contravene some other widely accepted norm or rule.

Some other truths are normative, I have claimed, in a different, stronger, *reason-implying* sense. Some examples are reason-implying moral truths, and truths about what we have non-moral reasons to believe, or to want or to do. Such truths, Non-Naturalists believe, are irreducibly normative. These truths cannot be explained in naturalistic terms. We cannot give some naturalistic description of certain acts, events, or states of affairs, and then say: 'That's *what it is* for some fact to give us a decisive reason,' or 'That's *what it is* for some act to be wrong.' Nor are these truths natural facts in the sense of being empirically discoverable. There could not be any empirical evidence for or against the belief that we have reasons to want to avoid future pain, or the belief that torturing people for our own amusement would be wrong.

As we shall see, there is one complication here. Some irreducibly normative *concepts* can refer to natural properties. But *claims* that use these concepts could not state natural facts. Nor could such facts be stated by any other irreducibly normative claims. When we have decisive reasons to do something, this fact couldn't be the same as some causal or psychological fact, such as the fact that this act would achieve one of our aims. And when certain acts are right, or wrong, these facts couldn't be the same as

certain natural facts, such as the facts that these acts would or would not minimize suffering, or would be acts of which most people would approve or disapprove.

2 Ontology

> We can next roughly distinguish between some views which differ *ontologically*, in the sense that they make different claims about what exists, or what is real. These views apply to all truths, not merely normative truths. Here are three such views:
>
> *Alethic Realism:* all true claims are made to be true by the way in which these claims correctly describe, or correspond to, how things are in some part of reality.
>
> *Naturalism about Reality*: The natural, spatio-temporal world is the whole of reality.
>
> *Alethic* Naturalism: All truths are about natural facts.

Some people are Alethic Naturalists because they are Alethic Realists who are also Naturalists about Reality. Some other Alethic Realists are not Naturalists. These people believe that certain claims are made to be true by being correct descriptions of how things are, not in the natural world, but in some other part of reality. In their beliefs about these truths, these people are *Metaphysical Non-Naturalists*.

Some other people reject Alethic Realism. Some of these people are what I call *Non-Realist Cognitivists*. We are Cognitivists but not Realists about some kind of claim if we believe that such claims can be true, but we deny

That these claims are made to be true by correctly describing, corresponding to, how things are in some part of reality.

This disagreement applies, for example, to arithmetical truths, such as the truth that there are infinitely many prime numbers. According to some Metaphysical Non-Naturalists, the numbers do not exist in the natural world, they exist in some other, non-spatio-temporal part of reality, such as what some people call *a Platonic realm*. Non-Realist Cognitivists make no such claim. True claims about numbers, these people believe, do not imply that numbers exist in some ontologically weighty sense either the natural world, or in some other non-spatio-temporal part of reality.

Similar claims apply to logical and modal truths. Two examples are the truths that some argument is valid, and that two plus two must equal four and could not possibly equal three or five. Validity is not a natural, empirically discoverable property, and when we make the modal claim that two plus two must equal four, we are not merely describing how things are in the actual world. There is no possible world in which two plus two would not equal four. If we are Non-Realist Cognitivists, we deny that such

logical and modal claims are made to be true by there being some part of reality which these claims correctly describe, or to which they correspond. If there is any dependence here, this dependence would go the other way. It would be reality that must correspond to these truths. Not even an omnipotent God could have made it false that two plus two equals four. We ought ... to accept some view of this kind.

Similar distinctions apply to views about normative truths; some norm-implying truths are ... [explained by] natural facts. We can explain in naturalistic terms how people can create certain rules or norms, and truths about these norms are empirically discoverable facts. There are some response-dependent normative truths which can be naturalistically explained and are empirically discoverable fats. When some meta-ethicists believe that all normative truths are of these kinds, that may be because these people are Alethic Realists who assume that the natural world is the whole of reality.

Some other truths are normative in a stronger, *reason-implying* sense. We cannot explain such truths in naturalistic terms, nor are such truths empirically discoverable. These non-empirical normative truths are I these ways like logical, mathematical, and modal truths. Since these are the truths that I shall be discussing, I shall use the word 'normative,' except when I say otherwise, in this stronger, reason-implying sense.

As in the case of other non-empirical turhts, our views about these truths can take two forms. *Metaphysical* Non-Naturalists believe that, when we make irreducibly normative claims, these claims imply that there exist some ontologically weighty non-natural entities or properties. Naturalists find such claims mysterious or incredible. Non-Realist Cognitivists deny that normative claims have any such ontological implications. On this view, normative claims are not made to be true by the way in which they correctly describe, or correspond to, how things are in some part of reality. Nagel, Scanlon, I, and others accept and defend one such view.

It might be objected that, in distinguishing these views, I have not explained what I mean by 'ontologically weight' or 'some part of reality'. But I use these phrases when describing views that I don't accept, and one of my objections to these views is the obscurity of their ontologival claims. When we consider concrete objects in the spatio-temporal world, such as starts, rocks, and people, we can make the clear and useful claim that these objects exist in an ontologically weighty sense. This is the sense in which we can deny that there are such things as ghosts, or phlogiston, or Cartesian Egos. We can also make such useful claims about the properties of concrete objects that can have causes and effects. Causal properties are of a distinctive ontological kind. But we cannot usefully make such claims when we consider entities and properties of some other kinds. When Platonists and Nominalists discuss mathematics, for example, they ask whether numbers and some other abstract entitities exist in some fundamental ontological sense, though these entitites do not exist in space or time. This question is not, I believe clear enough to be worth discussing.

When some ontologists discuss such questions, they claim that the word 'exist,' and the phrase 'there are,' have only one serious sense, which Quine calls 'the literal and basic sense.' We know what it is for stars to exist, these people claim, so we should be

able to understand the question whether, in this same sense, numbers exist. We ought, I argued, to reject this Single Sense View. Concrete objects and their causal properties both exist in what we can call a narrow, *actualist* sense. We can use the phrase 'there are' in another *possibilist* sense. We can truly claim, for example, that there was a possible palace designed by Christopher Wren which was never built, so that it never actually existed. This claim is not a contradiction, because the phrase 'there was' does not here mean 'there actually existed.' And we can often truly claim that there was something else that might have happened, or something else that we could have done. There is also a clear *non*-ontological sense in which there are many kinds of abstract entity. There are, for example, infinitely many prime numbers. But we cannot helpfully ask whether such abstract entitites exist in some different, ontologically weighty sense. Since I deny that this question is clear enough, I don't need to say more to explain my use of the phrase 'ontologically weighty.'

We can also use the words 'real' and 'reality' in different senses. In stating the view that I call Non-Realist Cognitivism, I have used the word 'reality' in a fairly clear, ontologically weighty sense. In this sense, merely possible objects, acts, or events are not part of reality, nor are abstract entities, such as valid arguments or prime numbers. We might instead use the word 'reality' in a wider sense, which implies that all truths are truths about reality. If that is how we use this word, the phrase 'about reality' adds nothing to the word 'true'. Nor could we say without self-contradiction that some true claims are not made to be true by the way in which they correctly describe, or correspond to, some part of reality. But if that is how we use the word 'reality,' we could restate Non-Realist Cognitivism in a different way. We could claim that, on this view, some non-empirical truths – such as logical, mathematical, and normative truths – do not raise any difficult ontological questions. Mathematicians, for example, should not fear that arithmetical claims might all be false, because there aren't any numbers. Nor should we fear that our non-empirical normative beliefs might all be false, because there are no non-natural properties of being right or wrong, or being good or bad, or being a normative reason.

Here is another way to sum up this view. Alethic Realists believe that

> (A) all true claims are made to be true by the way in which they correctly describe, or correspond to, how things are in some part of reality.

Metaphysical Naturalists believe that

> (B) the natural world is the whole of reality.

We cannot, I believe, defensibly combine these claims. If we use the word 'reality' in an ontologically weighty sense, and we accept (B), we ought to reject (A). We ought to believe that some true claims are not made to be true by corresponding to how things are either in the natural world or in some other part of reality. Some examples

are the kinds of non-empirical truth listed above. If instead we use the word 'reality' in a wider sense, which implies that all truths are truths about reality, we ought to reject (B). We ought to believe that there are some non-empirical truths that are not about the natural world. These truths, we can add, do not raise difficult ontological questions.

Non-empirical truths do, I believe, raise some difficult *philosophical* questions. Some of these questions are *epistemic*, in the sense that they are about whether and how we can justify our beliefs in these non-empirical truths. These truths may also raise some difficult *metaphysical* questions, such as questions about possibility and necessity. But these questions are not *ontological*, since they are not about whether certain entities or properties are *real*, or *exist*, in what some ontologists claim to be some single, deep, fundamental sense... .

When [Allan] Gibbard and Simon Blackburn discuss normative questions, they defend an original, subtle, and distinctive view to which Blackburn gave the name *Quasi-Realist Expressivism*. ... Gibbard and Blackburn are only *Quasi*-Realists – or *As If* Realists – in the sense that, though they believe that some normative claims are true, they reject both Normative Naturalism and Metaphysical Non-Naturalism. Blackburn calls his view 'anti-realist' because he denies that 'when we moralize we respond to, and describe, an independent aspect of reality.' Gibbard rejects what he calls the 'mysterious' idea that 'there is a normative realm distinct from the natural realm, and that we have ways to discern how things stand in that realm.' Gibbard and Blackburn are in one sense *Cognitivists*, since they believe that there are some normative truths. But they deny that normative claims are made to be true by corresponding to how things are in some part of *reality*. As we can more briefly say, Expressivist Quasi-Realism is another form of Non-Realist Cognitivism.

(vol. 3, 55–64)

/To see why this is the best way to understand Expressivism, see vol. 3, chapters 40, 45–47.]

[In this chapter, I focus instead on comparing my "... Non-Realist Cognitivism with Normative Naturalism. I shall first roughly describe some of the concepts that I shall use./

(vol. 3, 64)

3 Concepts and properties

> Of the words that we can use in stating our beliefs, some are names which merely *refer* to some person or thing. Two examples are 'Shakespeare' and 'Venus'. We use some other words and phrases to *describe* something. When two descriptive words or phrases mean the same, they express the same *concept*. The words 'new' and 'nuevo,' for example, both express the concept *new*. Some descriptive words and phrases refer to something by describing this thing. Two examples are 'the writer of Hamlet,' which refers to Shakespeare, and 'the lightest element,' which refers to hydrogen. There are also some partly descriptive names. One such name is 'the Morning Star,' which was once used to refer to the brightest starlike object that can sometimes be seen in the Eastern sky shortly before dawn. Another such name is 'the Evening Star,' which was once used to refer to the brightest starlike object that can sometimes be seen in the Western sky shortly after dusk.

These names are partly descriptive in the sense that we can easily describe the things to which they refer. Another example is the word 'water,' which is the English name of the transparent drinkable liquid that falls from the clouds as rain, flows in mountain streams, etc. Most names of people are not partly descriptive in this way. These distinctions are not sharp. Though we can tell people how to use the name 'red' by describing red as the colour of blood, to tell people how to use the names of particular shades, such as 'scarlet,' 'crimson,' or 'vermillion,' we may need to point to parts of a colour chart. As these examples show, some descriptive phrases or names refer, not to things, such as water or blood, but to the *properties* of things, such as the colour of blood. There are some distinctions here that are easy to overlook. Though the phrase 'the lightest element' refers to hydrogen, the similar phrase 'being the lightest element' refers, not to hydrogen, but to hydrogen's property of being the lightest element.

I use the word 'property' in the wide, non-legal sense in which any claim about something can be restated as a claim about this thing's properties. Instead of saying that the Sun is bright, or that some argument is valid, we can say that the Sun has the property of being bright, and that this argument has the property of being valid. Since this sense of the word 'property' adds nothing to the content of our claims, some people call it *pleonastic* or redundant. These remarks may suggest that this sense of the word 'property' is not worth using. But this pleonastic sense of 'property' can help us to explain the meaning of some claims, and to draw some important distinctions.

We can refer to such properties, for example, to explain the difference between two senses of the word 'is.' When we say that water is H_2O, we are using the 'is' of identity. We mean that water *is the same as* H_2O. When we say that the Sun is bright, we are not using the 'is' of identity, since we don't mean that the Sun is the same as being bright. We are using the 'is' of *predication*, since we mean that the Sun *has the property* of being bright.

Since this pleonastic sense of 'property' adds nothing to the content of our claims, our remarks about such properties have no ontological implications. Brightness and validity, though they are both pleonastic properties, differ ontologically in other ways.

These pleonastic properties we can also call *description-fitting* in the sense that they fit the descriptive words or phrases with which we refer to them. Because the word 'luminous' means 'radiates light,' the phrase 'being luminous' describes, and thereby refers to, the property of radiating light. Because the word 'trilateral' means 'having three straight sides,' this word describes and thereby refers to the property of having three straight sides. Such descriptive words or phrases also tell us *what it would be* for something to have some property. When something radiates light, that's *what it is* for this thing to be luminous, and having three straight sides is *what it is* to be trilateral. When some word or concept describes something, this word or concept *applies* to this thing. When we describe blood as red, for example, this use of 'red' applies to blood. Many descriptive words or phrases both apply to something and refer to this thing. The phrase 'the lightest element' both applies to and refers to hydrogen.

Some descriptive words and phrases, and the concepts they express, partly misdescribe the entity or property to which they apply or refer. That is true of the partly descriptive names 'the Evening Star' and 'the Morning Star.' Astronomers discovered that these names refer, not to different stars, but to the planet Venus. Another example is the concept of *an atom*. The word 'atom' originally meant, roughly, 'one of the smallest indivisible things of which physical objects are composed.' When physicists discovered that what they believed to be atoms are in fact divisible, since these things are composed of sub-atomic particles, they did not conclude that these things are not atoms. They revised their concept of *an atom* so that this complex concept ceased to include the concept of being indivisible. Though there can be some *mismatch* between the meaning of descriptive words, or the concepts these words express, and the entities or properties to which they refer, this mismatch cannot be great. Physicists could not have discovered that atoms are not things of which physical objects are composed. Nor could astronomers have discovered that the Evening Star and the Morning Star were not starlike objects in the sky, but optical illusions.

When there is nothing that sufficiently closely fits the meaning of some descriptive word or concept, we should claim that there is no such entity, or that nothing has the description-fitting property to which this word or concept refers. There have never been any witches, for example, because no one has ever had the property of being a witch. As these remarks imply, when we claim that some concept refers to some property in the description-fitting sense, we are not claiming that anything *has* this

property. No one has the property of having jumped over the Moon, and nothing could have the property of being a round square.

Since our words or phrases, and the concepts they express, refer to the properties that they describe, we might assume that words or phrases with quite different meanings must refer to different properties. That is not so. Different words or phrases may refer to the same property, which they accurately describe in different ways. As we shall see, it can be of great importance whether two such different descriptions refer to the same property.

The word 'property' is often used, not in the description-fitting sense, but in some narrower, ontologically weighty sense. In what I have called one such clear and useful sense, properties are the features of concrete objects or events which can have causes or effects. Being luminous is one such property, since light has causes and effects. But there are no such causal properties as those of being a prime number, a valid argument, or a normative reason.

Some people use the word 'property' in a third sense. The *extension* of any concept is everything to which this concept applies. The extension of the concept *red*, for example, is everything that is red, and the extension of the concept of *a prime number* is all of the prime numbers. Some concepts *necessarily* apply to something, in the sense that these concepts could not have failed to apply to this thing. The concept of *a prime number*, for example, necessarily applies to the number 7, since 7 could not have failed to be a prime number. Different concepts are *necessarily co-extensive* when these concepts necessarily apply to all and only the same things. Such concepts refer to the same property in what we call the *necessarily co-extensional* sense.

There is a distinction here which, as I have warned, we can easily overlook. Some concepts refer to something as the thing that has a certain property, and other similar concepts refer instead to this property. Consider first the concepts expressed by these phrases:

> the only even prime number,
>
> the positive square root of 4.

Each of these concepts refers, not to the number 2, but to one of the properties of this number. Since these concepts both necessarily apply only to the number 2, we can claim that

> (C) *being the only even prime number* and being the positive square root of 4 are the same property in the necessarily co-extensional sense.

We can also claim that:

> (D) these properties are different in the description-fitting sense. The concept of *being the only even prime number* does not describe, and thereby refer to, the property of *being the positive square root of 4*.

Suppose that some child doesn't understand the claim that 2 is the only even prime number. If this child's teacher used the word 'property' only in the co-extensional sense, this teacher might say: 'I told you yesterday what it is for some number to be the positive square root of 4. That is the same as being the only even prime number.' These remarks, though in one sense true, would be unhelpful. These two phrases describe and thereby refer to different properties. As we can more helpfully say, it's one thing to be the positive square root of 4, and a quite different thing to be the only even prime number.

Since (C) and (D) use the word 'property' in these different senses, these claims do not conflict. The description-fitting sense is more informative, by drawing distinctions which the co-extensional sense ignores. That is like the way in which some pairs of different geometrical shapes – such as the shapes of a sphere and a cube, or the shapes of a doughnut and a cup with one handle – are *topologically* the same. Just as topologists ignore many geometrical differences between different shapes, when we ask whether different descriptive concepts refer to the same property in the necessarily co-extensional sense, we ask only *which* are the things to which these concepts apply, ignoring all of the differences between the ways in which these concepts describe these things.

For another example, we can suppose that our concept of a *human being* makes it a necessary truth that all human beings are both conceived and later die. We could then claim that *being a human who was conceived* and *being a human who later dies* are the same property in this co-extensional sense. But these properties are not the same in the more informative description-fitting sense. Being a human who was conceived is in this sense different from being a human who later dies. This is how we can rationally be glad that we were conceived but regret that we shall die.

(vol. 3, 65–70)

4 Against normative naturalism

> I shall now discuss some arguments for and against Normative Naturalism, which I shall try to state more clearly than I did in my *Volume Two*. I shall also correct some mistakes, and add some further claims. There are at least two kinds of normative truths. Some truths are normative in the sense that they are about widely accepted rules or norms. As I have said, these normative truths are also natural facts in the sense that they are empirically discoverable. We can discover, for example, whether certain acts are, in some community, illegal, or contravene other widely accepted norms. I shall mainly discuss truths that are normative in a stronger, reason-implying sense. These truths are *irreducibly* normative in the sense that they cannot be restated in non-normative naturalistic terms. These non-empirical normative truths are in these ways like logical, mathematical, and modal truths.

Normative Naturalists reject some of these claims. Some *Analytical* Naturalists deny that there are any such irreducibly normative concepts and claims. But this view is clearly false. Some normative concepts, such as the concepts *wrong* and *a decisive reason*, cannot be correctly explained in non-normative, naturalistic terms. According to *Non-Analytical Naturalists*, though some of our concepts and claims are irreducibly normative, these concepts refer to natural properties, and these claims, when they are true, state natural facts. Such views are much more plausible.

In considering the argument for and against such views, I shall apply them to one of the simplest moral theories, Hedonistic Act Utilitarianism. I shall here state this view as

> HAU: Acts are right if and only if, or *just when*, these acts minimize the total sum of suffering minus happiness.

Acts that minimize this total sum – or, for short, that minimize suffering – can also be described as maximizing the total sum of happiness minus suffering. My less familiar statement of HAU better expresses, I believe, what makes this view plausible. We need not decide whether HAU is true, since most of the claims and arguments that we shall be considering could be applied to other moral views, and to other reason-implying normative claims.

If some view like HAU were true, this view would not merely happen to be true, but would be a necessary truth, which would be true in all possible worlds. To give

a more widely accepted view, there could not be any world in which it would right for some conscious rational beings to torture others for their own amusement. There are some other, less fundamental normative truths, which would be true in only some possible worlds.

(vol. 3, 70–71)

a The co-extensiveness argument

> When some Normative Naturalists defend their view, they appeal to the necessity of some normative truths. These people might argue:
>
> If HAU were true, the concepts *right* and *minimizes suffering* would necessarily apply to all and only the same acts.

When two concepts are necessarily co-extensive, these concepts refer to the same property.

Therefore

> If HAU were true, the normative property of *being right* would be the same as the naturalistic property of *being an act that minimizes suffering.*

When I earlier discussed this *Co-Extensiveness Argument*, I rejected its second premise. That was a mistake. I should have admitted that the phrase 'the same property' can always be used in the necessarily co-extensional sense. I should then have denied that this argument supports Naturalism. Non-Naturalists could reply:

> (E) Even if the concepts *right* and *minimizes suffering* referred to the same property in this co-extensional sense, these concepts would refer to different properties in the description-fitting sense. If HAU were true, acts that have the natural property of minimizing suffering would also have the different, irreducibly normative property of being right.

For Naturalists to defend their view, they must reject (E), claiming instead that

> (F) if HAU were true, *being right* and *being an act that minimizes suffering* would be the same property in the description-fitting sense.

This claim, I believe, could not be true. The concept of *being an act that minimizes suffering* does not describe, and thereby refer to, the property of *being right.* Nor could this normative property be described, and thereby referred to, by any other naturalistic concept, such as the concept of *being an act of which most people would approve.*

(vol. 3, 71–72)

b The normativity objection

> I have just stated the simplest and most straightforward objection to Normative Naturalism. According to this
>
> *Normativity Objection:* Irreducibly normative, reason-implying claims could not, if they were true, state normative facts that were also natural facts.

These two kinds of fact are, I believe, in two different, non-overlapping categories. There are many such different categories. It could not, for example, be a physical or legal fact that 7 * 8 = 56, nor could it be a legal or mathematical fact that galaxies rotate, nor could it be a physical or mathematical fact that perjury is a crime. As these examples suggest, it would not be surprising if no natural facts, such as causal, psychological, or sociological facts, could also be irreducibly normative, reason-implying facts.

Some act is right, in the sense that I shall use, if this act is what we ought morally to do, because every other possible act would be wrong. We can use the word 'wrong' in some definable senses, such as those expressed by the phrases 'blameworthy,' unjustifiable to others,' 'something that we have morally decisive reasons not to do,' and 'an act that gives the agent reasons to feel remorse, and gives other people reasons for indignation.'

According to Normative Naturalists, such definable senses of 'right' and 'wrong' might refer to some natural property. That could not, I believe, be true of the senses and concepts that I have just mentioned. These senses and concepts refer to the properties they describe. The concept *blameworthy*, for example, refers to the property of being blameworthy, and the concept *unjustifiable to others* refers to the property of being unjustifiable to others. These concepts do not describe, and thereby refer to, any natural property, such as the properties of being an act that would minimize suffering, or an act of which most people would disapprove. Similar claims apply to most other definable normative concepts. There are some important exceptions, however, to which I shall return.

When we discuss some simpler normative concepts, the Normativity Objection must be stated in a different way. We cannot helpfully define the concept of *a normative reason*, or the non-moral concept *ought* which implies that we have decisive reasons to have some belief or desire, or to act in some way. Nor can we define the sense of 'wrong' that we can also express with words like 'impermissible' or 'mustn't-be-done.' These, I believe, are the most important normative concepts, which can be used to state the most important normative beliefs. These indefinable concepts do not describe some property, so we cannot claim that they refer to the properties that they describe. These concept are in this way like non-descriptive names, so we must explain in some other way which are the properties to which these concepts refer.

Since these indefinable concepts are like names, Naturalists might say that, if it were true that acts are right just when they minimize suffering we could claim that minimizing suffering is the natural property to which we refer by using the name 'right.'

When words or names are indefinable, however, we may know the *kind* of thing or property to which they refer. We cannot helpfully define some other fundamental concepts such as *time, space, necessary*, and *possible*, but we understand these concepts well enough to be able to reject most mistaken claims about the entities or properties to which these concepts refer. Though it is hard to explain what time and space are, we know that there are countless things that time and space *couldn't* be. It is similarly true, I believe, that these indefinable normative concepts couldn't be names which referred to natural properties. But this belief is much less obviously true, which is why we need to consider arguments for and against this belief.

To tell people how to use names like 'Scarlet,' or 'Sirius,' we can point to a colour chart, or to a star. Indefinable normative concepts are harder to explain, partly because we can't point to the properties to which these concepts refer. But we may get people to think thoughts that involve these concepts. Discussing an imagined case that I called *Burning Hotel*, I asked my readers to suppose that they will soon die unless they jump into some canal. I wrote:

Since your life is worth living, it is clear that

> (1) you ought to jump.

This fact, some Naturalists claim, is the same as the fact that

> (2) jumping would do most to fulfil your present fully informed desires, or is what, if you deliberated in certain naturalistically describable ways, you would choose to do.

Given the difference between the meanings of claims like (1) and (2), such claims could not, I believe, state the same fact. Suppose that you are in the top storey of your hotel, and you are terrified of heights. You know that, unless you jump, you will soon be overcome by smoke. You might then believe, and tell yourself, that you have *decisive reasons* to jump, that you *should, ought to*, and *must* jump, and that if you don't jump you would be making a *terrible mistake*. If these normative beliefs were true, these truths could not possibly be the same as, or consist in, some merely natural fact, such as the causal and psychological facts stated by claims like (2).

This objection to Naturalism, we can add, need not assume that, as Non-Naturalists believe, there are some irreducibly normative truths or facts. Many Non-Cognitivists and Error Theorists also believe that some normative claims are in a separate, distinctive category, so that these claims could not state natural facts. Some of these people add that, since all facts are natural, there are no normative facts.

(vol. 3, 72–75)

c Scientific analogies

> This Normativity Objection fails to convince some Non-Analytical Naturalists. These people agree that some normative concepts and claims are quite different from naturalistic concepts and claims. But these people argue that these irreducibly normative concepts might refer to natural properties, and these irreducibly normative claims might state natural facts. Some of these people appeal to the discoveries that
>
> (A) heat is molecular kinetic energy, and that
>
> (B) water is H_2O.

These scientific discoveries were not implied by the pre-scientific meanings of the words 'heat' and 'water.' It might be similarly true, these people claim, that some normative and naturalistic concepts refer to the same property. We might then be able to use these concepts to state normative truths that were also natural facts.

These claims are plausible, and have been well explained and defended by many people. But when we look more closely at these scientific analogies, we find, I believe, that they do not support Naturalism.

Some Naturalists claim that

> (C) as these scientific analogies show, truths about the identity of properties may not match, or closely depend upon, the concepts with which we refer to these properties.

That is not, I believe, what these analogies show. Most descriptive words or phrases, and partly descriptive names, fairly accurately describe the entities or properties to which they refer. There can, I have said, be some mismatch between these descriptions and these entities or properties. That was true when astronomers discovered that the Evening Star was not a star but a planet, and when physicists discovered that atoms are not – as the meaning of the word 'atom' implied – indivisible. But such a mismatch could not be great. Though the truth of both (A) and (B) had to be discovered, there was no mismatch between the meaning of the words 'heat' and 'water' and the things to which these words refer. In its relevant, pre-scientific sense, the word 'heat' mean, roughly:

> the property that can have certain effects, such as causing us to feel certain sensations, melting solids, turning liquids into gases, etc.

The word 'heat' *does* refer to the property that can have these effects. This property, scientists discovered, can also be truly described in a different way, as the property of having molecules that move energetically. We can therefore claim that

(D) *being hot* and *having molecular kinetic energy* are the same property in the description-fitting sense.

Similar remarks apply to the fact that water is H_2O. In its pre-scientific sense, 'water' is a partly descriptive name which refers to the liquid that is transparent, drinkable, falls from the sky as rain, etc.

The liquid that can be truly so described, scientists discovered, can also be truly described as being composed of molecules of H_2O. We can therefore claim that

(E) *being water* and *being composed of* H_2O are the same property in the description-fitting sense.

We should agree that, as many Naturalists point out, the truths stated by (D) and (E) were not implied by the meanings of the words 'heat,' 'molecular kinetic energy,' 'water,' and 'H_2O.' But these facts do not count against the view that these words, and the concepts they express, refer to the properties that they describe. These cases show only that, when two different concepts correctly describe and thereby refer to the same property, this fact may not be directly implied by these concepts. We may have to discover this fact, as scientists did with (D) and (E), or come to know this fact in some other way.

Here is another way to make this point. To defend the claim that heat is a molecular kinetic energy, we must use the word 'heat' to refer to the property that can have certain effects, such as causing us to feel certain sensations. If instead we used the word 'heat' in a different pro-scientific sense, with which we intended to refer to sensations of heat rather than to the property that causes these sensations, scientists could not have discovered that these sensations were the same as molecular kinetic energy.

Some Naturalists, I have said, claim that

(C) truths about the identity of properties may not match, or closely depend upon, the concepts with which we refer to these properties.

As I have just argued, however, Non-Analytical Naturalists cannot defend their view by appealing to (C). These people believe that some concepts and claims, though they are irreducibly normative, might refer to natural properties and state natural facts. Since these Naturalists cannot appeal to (C), they would first have to explain how it might be true that

(F) some irreducibly normative concepts refer to natural properties.

They would then have to show that

(G) we can use the normative concepts to make irreducibly normative claims which, if they were true, would state natural facts

Most definable normative concepts could not, I have claimed, refer to natural properties. But there are some important exceptions, to which we can now turn.

The scientific analogies are in one way helpful here, since they suggest how (F) might be true. The pre-scientific concept of heat has what I called a *gap that is waiting to be filled*. The concept refers to the property, *whichever it is*, that can have certain effects, such as causing certain sensations, melting solids, turning liquids into gases, etc.

This property, scientists discovered, is that of having molecules that move energetically. There are, I claimed, some similar normative concepts. One example is the property of the natural property, *whichever it is*, that makes acts right.

This concept, we might say, is the concept of the *right-maker*. Since this complex concept includes the concept *right*, the concept is irreducibly normative. Naturalists might claim that

> (H) though this concept is irreducibly normative, it might refer to a natural property. If HAU were true, being an act that minimizes suffering would be the right-maker, in the sense of being the natural property that makes acts right. We could then conclude that minimizing suffering is the same as being right.

This claim, I argued, cannot be true. Non-Naturalists could reply that

> (I) If HAU were true, being an act that minimizes suffering would be the natural property that made acts have the different, normative property of being right.

In my Volume Two, I suggested how Naturalists might reject this reply. When some fact about some act *makes* this act right, this fact doesn't cause this act to be right. *Making right* is a non-causal relation. There are other ways in which, when something has some property, this fact may *non-causally make* this thing have some property. We have been discussing two examples. When some liquid is composed of H_2O, this fact *makes* this liquid water but it doesn't *cause* this liquid to be water. When the molecules in some object move energetically, this fact *makes* this object hot but doesn't *cause* this object to be hot. Some Naturalists might claim:

> As these cases also show, the relation of *non-causal making* implies *being the same as*. When some object has molecular kinetic energy, this fact both makes this object hot and is the same as this object's being hot. When some liquid is composed of H20, this fact both makes this liquid water and is the same as this liquid's being water. It is similarly true that, if there is some natural property that makes acts right, this natural property would be the same as the property of being right.

I rejected these claims. There are, I claimed, several kinds of non-causal making. Having a child, for example, non-causally makes us a parent, but this analytic truth is unlike the scientific truth that having molecular kinetic energy non-causally makes some object hot. I then claimed that if there is some natural property which is the right-maker, in the sense of being the property that makes acts right, this non-causal making relation would be of a third, distinctive kind. This third relation does not imply being the same as. We should instead believe that, when acts have this natural property, that would non-causally make these acts have the different property of being right.

These Naturalists might reply:

Your claims are mere assertions. You agree that

> (1) having molecular kinetic energy both non-causally makes an object hot and is the same as being hot.

We believe that, if HAU were true, we could similarly claim that

> (K) being an act that minimizes suffering both non-causally makes an act right and is the same as being right.

Why do you deny that (K) could be true? Where is the disanalogy?

This reply is, I concede, plausible. Given the similarity between (J) and (K), it may seem dogmatic to declare that being an act that minimizes suffering couldn't be the same as being right.

There is, however, a better objection to this argument for Naturalism, which I earlier overlooked. (J) and (K) are not, I believe, relevantly similar claims. The wording of these claims suggests that what corresponds to *being hot* is *being right*. But that is not so. Heat is *the property that can have certain effects* – such as causing certain sensations, melting solids, etc. What corresponds to heat is not *the property of being right*, but *the property that makes acts right*. What is relevantly similar to (J) is not (K) but

> (L) being an act that minimizes suffering both non-causally makes an act have the property that makes acts right, and is the same as having the property that makes acts right.

I should have claimed that, unlike (K), (L) might be true. (L) is like the true claim that

> (J) having molecular kinetic energy both non-causally makes an object hot and is the same as being hot.

The analogy between such normative and naturalistic truths is therefore closer than I earlier claimed it to be.

This closer analogy, however, still fails to support Naturalism. Non-Naturalists could accept that, if HAU were true, being an act that minimizes suffering would be the same as *having the property that makes acts right*. But Non-Naturalists could deny that, if HAU were true, being an act that minimizes suffering would be the same as *being right*.

Naturalists might reject these claims. They might say:

> (M) As these scientific analogies show, having the property that makes acts right is the same as being right.

I shall now try to explain more clearly why I believe that (M) is false. We can first note that

> (N) nothing can be the same as one of its properties.

This truth is obvious when we compare some concrete object with this object's properties. No one would confuse the Sun with the Sun's property of being bright. It can be easy, however, to confuse some property with some of the properties *of* this property. We should therefore add that

> (O) no property can be the same as any of its higher-order properties.

One such truth is:

> (P) When some property has some effect, this property can't be the same as its higher-order property of being the property that has this effect.

We can add that

> (Q) the property *that has* some effect cannot be the same as the property *of having* this effect, nor can it be the same as *this effect*.

Stated so abstractly, these points can be slippery, and hard to understand. So we can return to our example. The Sun's brightness, during a cloudless night, makes the Moon shine. We can claim:

> (R) Just as the Sun couldn't be the same as the Sun's property of being bright, the Sun's brightness couldn't be the same as any of this property's higher-order properties, such as the property of being the property that makes the Moon shine. Nor could the Sun's brightness be the same as the property of being made to shine, or of shining.

Return now to the Naturalist's claim that

(M) having the property that makes acts right is the same as being right.

Though *making right* is not a causal relation, similar remarks apply. We can claim that

> (S) just as the property that makes the Moon shine couldn't be the same as this property's higher-order property *of being* the property that makes the Moon shine, the natural property that makes acts right couldn't be the same as this property's higher-order property of *being* the property that makes acts right. Nor could the property that makes acts right be the same as the property of *being made* to be right, or the property of *being right*.

These claims show, I believe, that (M) is false. We should conclude that

> (T) the natural property that makes acts right couldn't be the same as the normative property of being right.

Some Naturalists might reject these claims, by returning to the claim that

> (U) the relation of *non-causal making* always implies *being the same as*.

As I earlier claimed, however, (U) is false. We can first note that, even when the relation of *non-causal making* implies the relation of *being the same as*, these are different relations. Some relations, including *being the same as*, are *symmetrical*, in the sense that they hold in both directions, and can be reversed. If A is the same as B, B must be the same as A. The relation of *non-causal making* is, in contrast, *asymmetrical*. Though having molecular kinetic energy makes something hot, being hot doesn't make something have molecular kinetic energy. Though having a child makes us a parent, being a parent doesn't make us someone who has a child. We can similarly claim that, if some natural property were the property that made acts right, having this natural property would make acts right, but being right would not make acts have this natural property.

We can next point out that, as well as being asymmetrical, the relation of *non-causal making* often holds between different properties. If we drive dangerously, that makes our act illegal, but driving dangerously isn't the same as being illegal. If some act is illegal, that makes this act punishable, but being illegal isn't the same as being punishable. If we were killed by a meteorite, that would make us unlucky, but being killed by a meteorite isn't the same as being unlucky. If Mozart had lived longer, and written ten more operas, that would have made him an even greater composer, but Mozart's writing of ten more operas would not have been the same as his being an even greater composer. We can similarly claim that

> (V) even if minimizing suffering were the property that makes acts right, minimizing suffering couldn't be the same as being right.

Similar claims apply to any other natural property. No such property could both be the property *that makes* acts right and be the property *of being* right.

These claims have been about one particular normative concept, which is the concept of non-causally making acts right. Even If Naturalists now accept (V), they might claim that

> (W) there may be other irreducibly normative concepts that might refer to natural properties.

They might then claim that

> (X) claims that used these other normative concepts might, if they were true, state natural facts.

(W), I believe, is true. Two such concepts might be those of the natural property that has the greatest moral importance, and the natural property whose being had by people we have the strongest reasons to regret.

These normative concepts might both refer to the natural property of having intense and prolonged suffering.

Though (W) is true, however, (X) is false. If prolonged and intense suffering is the natural property that has the greatest moral importance, and is the property that we have the strongest reasons to regret, these would be normative truths. Similar claims apply to all such concepts and claims. If these normative concepts referred to some natural property, they would refer to this property *as* the natural property that has some other, irreducibly normative property. Since these concepts would refer to some natural property as the property that has some higher-order normative property, claims that used these concepts could not state natural facts.

Since I have been discussing some particular ways in which Naturalists might reply to the Normativity Objection, and I have made some complicated claims, it may be worth returning briefly to some of the most important normative concepts, and to the simplest version of the Normativity Objection. Suppose again that you are in the top storey of my imagined *Burning Hotel*, and you will soon die unless you jump into some canal. You tell yourself that you have *decisive reasons* to jump, that you *should, ought to*, and *must* jump, and that if you don't jump you would be making a *terrible mistake*. If these normative beliefs were true, these truths could not be the same as, or consist in, some natural empirically discoverable facts, such as the fact that jumping into the canal would fulfil your present desires, or is what, after informed deliberation, you would choose to do. These truths would be in quite different, non-overlapping categories.

There are, I have said, many such categories. Though the Normativity Objection appeals to the distinctive nature of irreducibly normative truths, this objection is merely one example of many other, similar claims. There are many such non-overlapping categories. Physical facts, for example, could not be the same as logical, legal, musical,

grammatical, exegetical, or mythological facts. Nor could any of these other facts be in two of these different categories. As these examples suggest, it would not be surprising if, as I believe I have now shown, non-empirical normative truths could not be the same as any naturalistically explainable and empirically discoverable facts. These answers to the Normativity Objection, I conclude, fail.

(vol. 3, 75–84)

d The triviality objection

Since some Naturalists were not persuaded by the Normativity Objection, I gave some others. According to all Non-Analytical Naturalists, irreducibly normative concepts and claims might refer to natural properties and state natural facts. Such views take two forms. According to

Hard Naturalists: Since all facts are natural, we don't need to use such normative concepts or make such normative claims.

[Frank] Jackson, for example, writes that, when we have reported the facts in naturalistic descriptive terms,

there is nothing more 'there' ... There is no 'extra' feature that the ethical terms are fastening onto, and we could in principle say it all in descriptive language.

According to

Soft Naturalists: We do need to use such normative concepts and to make such normative claims. Though true normative claims could state only natural facts, having true normative beliefs about these facts would help us to make good decisions and to act well.

This second view, I argued, could not be true. Irreducibly normative claims could not, I believe, state natural facts. But if – impossibly – these claims did state such facts, these facts would be normatively trivial, since they could not give us positive substantive normative information. I called this argument the *Triviality Objection.* This name is in one way misleading. If there were no irreducibly normative truths, this fact would not be trivial. I should have said only that, if there were no such truths, our normative beliefs could not help us to make good decisions and to act well.

Soft Naturalism is, in one way, hard to consider, because it is hard to assess counterfactuals with antecedents that could not possibly be true. It would be hard, for example, to assess the claim that, if you were an ant, you would do what ants do, or that, if no one had any reason to want to avoid agony, torture would not be wrong. But we can try to suppose that irreducibly normative claims did state natural facts, and ask whether such truths might support Soft Naturalism. For example, we can try to suppose that

(A) being an act that minimizes suffering is the same as being right.

If, impossibly, (A) were true, this claim might seem to give us positive substantive normative information. But that is not so. (A) would not tell us that, when some act minimizes suffering, this fact would make this act have the different normative property of being right. (A) would tell us that there is *no* such *different* property. Though (A) would give us substantive normative information, this information would be *negative*. If there was no such different property as that of being right, we would have wasted our time whenever we tried to decide which acts were right.

Soft Naturalists might reply, if (A) were true, this claim would *indirectly* give us positive normative information. If we learnt that being right was the same as having the natural property of minimizing suffering, this fact might indirectly tell us how this natural property was related to one or more other, normative properties. I considered various suggestions of this kind; but none, I argued, could succeed.

We can next consider the Naturalists who believe that, if it were true that acts are right just when they minimize suffering we could claim that

> (B) minimizing suffering is the natural property to which we refer by using the name 'right.'

This claim couldn't state a positive substantive normative truth. Knowing how we use some name couldn't help us to make good decisions and to act well.

Some other Naturalists claim that, by considering the complex role that the word 'right' plays in our moral thinking, we might be able to show that the property of being right consists in having any of several natural properties. This conclusion, these people might claim, would give us positive substantive normative information. But that is not, I believe, trut. One such view might claim that

> (C) when some act would be right, that is the same as this act's being either the saving of someone's life, or the keeping of a promise, or the paying of a debt ... and so on.

Most of us believe that

> (D) if some act would save someone's life, this act would be right.

According to these Naturalists, the fact stated by (D) could be restated as

> v if some act would save someone's life, this act would be either the saving of someone's life, or the keeping of a promise, or the paying of a debt ... and so on.

No one could doubt the truth of (E). But this truth would not give us any positive substantive normative information. Soft Naturalists could not defensibly claim that, if we believed some truth like (E), that would help us to make good decisions and to act well.

There are some other versions of Naturalism which appeal to the complex role that the concepts *right* and *wrong* play in our moral thinking. These are the most plausible versions of Naturalism. But the Naturalists who defend these views could not, I believe answer the objections that I have described above.

(vol. 3, 84–86)

5 Railton's agreement

> In his remarkably constructive and, to me, exhilarating paper, 'Two Sides of the Meta-Ethical Mountain,' Peter Railton ... [writes]:
>
> (vol. 3, 91)
>
> Soft Naturalists ... can accept *non-natural* properties in a nominal or linguistic – and to that extent *non-ontological* – sense. Soft Naturalists also allow us to talk meaningfully and truthfully in terms of normative concepts – making reasons claims in which *normative predicates* figure. True reasons claims can be called *normative facts* by the Soft Naturalist in one familiar sense of the term – a *fact* is the content of a true statement or proposition.
>
> Railton also writes:
>
> Naturalists can tolerate *linguistic* or *nominal* properties ... [and] the conveyance of certain kinds of information... . by citing such properties, so long as this does not involve adding anything to their *ontologies*... . Parfit claims that non-natural properties exist only in a *non-ontological sense*... . Reflecting on all this, I'm not sure how strenuously a Soft Naturalist should object, if at all, to 'non-ontological' non-natural properties or to the ... non-natural facts attributing them. Soft Naturalists surely object to Platonistic Non-Naturalism, complete with an ontic conception of non-natural properties ... but Parfit's Non-Naturalism is not of this kind ..." (vol. 3, 100)

Though Railton rejects Metaphysical Non-Naturalism, his objections do not apply, he writes, to the non-ontological normative claims made by those whom I now call Non-Realist Cognitivists. Some examples are claims about normative reasons, and about what we ought to do in the decisive reason-implying sense.

> (vol. 3, 100–101)

6 Gibbard's agreement

> Gibbard ... suggests that his Expressivist view should take ... Non-Realist Cognitivst form. There are, I have claimed, some truths which are non-natural, in the sense that they are not empirically discoverable, and non-ontological in the sense that they raise no difficult ontological questions. One example is the truth that there is a valid proof of some mathematical theorem. If there is such a proof, I claimed, this would not be an empirically discoverable fact about the natural world, nor is validity an ontologically weighty property. Gibbard calls it 'difficult and puzzling what to say about mathematics.' He suspends judgment on the question whether there are any such non-ontological properties.

In his latest book, however, Gibbard makes some striking positive claims. Traditional versions of Expressivism and Non-Naturalist Cognitivism, Gibbard writes, were 'far apart on many issues,' but these theories have made progress in ways that bring them closer together. In the best versions of these theories, Gibbard suggests, Quasi-Realist Expressivists would claim that there are some true irreducibly normative beliefs, and Non-Naturalists would drop their claim that these normative beliefs are about ontologically weighty non-natural properties. These theories, Gibbard writes, might then 'coincide in all their theses.'

> (vol. 3, 182)

Section 2: Some substantive matters

26. Non-requiring reasons

1 Introduction

Reasons for action, it is often said, are considerations that normatively speak in favor of an action (Scanlon 1998, p. 19, 2014, p. 30). So put, reasons can sound rather friendly. They offer "normative support;" they "rationalize an action;" they render it "rationally (morally, prudentially) eligible." In a classical approach, though, the way practical reasons do all of this is by way of issuing a deontic directive. Simply put, when we say that a reason speaks in favor of an option, we are saying it would be wrong not to follow the lead of that reason unless there were some countervailing justification not to. Depending on the type of reason, the wrong in question may be the wrong of immorality, or imprudence, or just the generic wrong of practical reason. And of course, countervailing justification there may well be. Still, in this view, practical reasons necessarily put one in danger of going wrong. Reasons for actions are normative entities inherently on their way toward being all things considered deontic oughts.

Defended explicitly by theorists as diverse as Robert Audi (1997) and Shelley Kagan (1989), the view is also sufficiently orthodox as to be tacitly assumed in many discussions.[1] Indeed, as Jonathan Dancy puts it, for many, this picture of how practical reasons function is often built into the very definition of a practical reason, with reasons defined as considerations it is "wrong not to act on in the absence of any opposition" (2004, p. 92).

Against this view, an increasing number of theorists have argued that reasons are not always in the requiring business. Endorsing rich and meaningful latitude in the lives of agents, they defend the existence of non-requiring or optional reasons. Thus Joshua Gert argues for the existence of "purely justifying reasons," "which can be very strong rational justifiers but which do not rationally require at all" (2004, p. 23). T.M. Scanlon defends the existence of "strongly optional" reasons, which "render an action rationally eligible without making it rationally required in the absence of some countervailing reason" (2014, p. 107). In the moral realm, Terrence Horgan and Michael Timmons defend "merit-conferring reasons," whose normativity involves "favoring but not requiring" (2010, p. 49). In the non-moral realm, Jonathan Dancy defends the existence of "enticing" reasons, which serve to make "an option attractive rather than

[1] As Gert says in a footnote (p. 19), "This view is so widespread that many theorists do not seem to recognize that there is a position opposed to it. As a result, it is not often clearly stated." Nevertheless, for relatively clear endorsements, see Darwall (1983, pp. 19, 54); Korsgaard (1996, pp. 225–226); Audi (1997, pp. 146–147); Scanlon (1998, pp. 18–23); Copp (1995, p. 42); Velleman (1996, p. 705ff); Edgley (1965, pp. 182–188). We would add Sergio Tannenbaum and Shelley Kagan.

demanded, required, or right" (2004, p. 99). Patricia Greenspan defends the existence of "purely positive reasons," which "do not compel, but instead are optional, rendering an option eligible for choice, or justifying it, without requiring it" (2005, p. 389). The list continues.[2]

We ourselves are enthusiastic members of this movement (Little 2013; Little and Macnamara 2017). That said, discussions here can be confusing, for in fact there are two very different types of reasons – with very different ways of underwriting or securing latitude – that have been advanced under the rubric of non-requiring reasons. One camp, led most prominently by Joshua Gert but including, for instance, Patricia Greenspan and Douglas Portmore (2012), defends the existence of reasons that play what we will call a permissibility-conferring role. These are reasons that function to defuse or neutralize, by a certain degree, the force of requiring reasons. Another camp, led most prominently by Jonathan Dancy, but including, for instance, Horgans and Timmons and Fiona Woollard (2016), defends the existence of reasons that play what we will call a commendatory role. These are reasons that function to normatively endorse an action without placing one in need of justification to decline. Compatible claims, they are also conceptually distinct. While one could endorse both kinds of reasons, one need not, and the arguments for one are not arguments for the other. Both represent important, distinct challenges to the standard view of practical reasons as universally and exclusively requiring.

In this chapter, we lay out the two camps – their core claims, animating motivations, and, in the final section, why they are sometimes confused for one another. We begin by describing the standard view both aim to challenge.

[2] See, for instance, Kauppinen's "evaluative reasons" (2015) and Kolodny's "non-insistent reasons" (2003).

2 The standard view

On one classic conception, reasons for action are deontic directives. If one faces a reason in favor of an action, then one would be wrong not to follow its lead absent sufficient countervailing justification. Depending on the type of reason, the wrong in question may be the wrong of immorality, or imprudence, or just the generic wrong of practical reason. And, of course, countervailing justification one may well have. Still, to understand something as a practical reason necessarily brings with it a kind of deontic vulnerability: one who faces such a reason now stands in need of adequate justification to do other than it directs, on pain of going wrong.

In this view, the normative force of a practical reason can be described as a kind of requiring. Like a demand, reasons place us in need of a certain amount and type of justification not to follow their direction. They favor an action, in essence, by saying it would be pro tanto (insofar as this consideration goes) wrong *not* to do it. If reasons for action sound friendly, then, in fact, they are something of a normative stick. Reasons for actions are normative entities inherently on their way toward being all-things-considered deontic oughts.

In addition to those who explicitly defend the claim, it is also implied by certain claims that many put forward as supposedly neutral depictions of reasons for action. For instance, it is often said that practical reasons are, by their very nature, governed by a principle of "motivational internalism." According to this principle, to believe that p is an undefeated reason to phi necessarily entails that one will be motivated to do as it directs in the absence of weak will or deliberative error.[1] This, note, just *is* to state that all reasons carry requiring force. An undefeated requirement is wrong not to follow, and wrong is a marker of what a fully functioning agent – one devoid of weak will or confusion – does not do.

Core to this standard view is what we might call a "monistic theory" about how reasons contribute to the rational or moral status of actions. In this view, the normative function of a practical reason to phi is to issue a pro tanto requirement to phi, displacing by some measure of strength the permissibility of not phi'ing, and this is the *only* normative function reasons for action can play. When determining whether a given action is supported by reasons, what we are doing is balancing, comparing, or otherwise adjudicating the strength, direction, and specificity with which the relevant reasons variously require. To assess reasons' contribution to the status of an action is to adjudicate the deontic push that each brings.

[1] For an example of motivational internalism applied to moral reasons, see Smith 1994.

As many have pointed out, this raises questions about the possibility for meaningful latitude. Since many who endorse the standard view do regard latitude as an important aspect of our lives as agents, various strategies have been advanced for recovering latitude within its strictures. Raz defends the existence of widespread latitude in everyday life by positing the existence of widespread incommensurability – considerations that are neither better than, less than, nor equal to one another.[2] The idea is as follows. When requiring reasons compete with and oppose one another, they provide a counter-balance to each other's deontic challenge: their presence provides a measure of countervailing justification for not following the others' direction. When a dominant reason emerges from the competition – when one is stronger than the rest – there is obviously no latitude, for one must obey the winner's command. However, latitude is possible when the reasons in question are incommensurable. Each member of the set suffices to justify not following the other, but since none is defeated, it is acceptable to act on any in the set.

Others have defended the existence of latitude in the moral realm by positing normative insulation between personal and moral reasons. For instance, Sergio Tannenbaum (2007) accounts for the possibility of supererogatory actions (actions that are morally good but not required) by arguing that moral and prudential reasons form dual perspectives that cannot be judged against one another from any unified perspective. Others defend specific accounts of vagueness to expand the cases in which competing reasons do not dominate one another. What all such strategies have in common is the idea that latitude is enjoyed only where the sets of competing requirements we face have no victor.

While there is a lively literature challenging the specifics of these strategies,[3] another set of theorists argues that the problem goes deeper than details. The source of the problem, they argue, in effect, is with the functional monism advanced by the standard view. Rich and meaningful latitude exists, they believe, not just because competing requirements sometimes admit of no winner, but because requiring is not all that practical reasons do.

[2] Raz 2000: chapters 3 and 5; see also 1989: chapter 13. Raz's account of *moral* latitude is different (1975). It appeals to a concept of exclusionary permissions, which can be seen as an early precursor of permissibility-conferring reasons. For Gert's discussion of Raz's exclusionary permissions, see pp. 106–109.

[3] Famous objections to widespread incommensurablility, for instance, can be found in Chang 1997; Gert 2004, pp. 102–105.

3 Permissibility-conferring reasons

The first group posits what we'll call permissibility-conferring reasons. If requiring reasons place one in need of a certain degree and kind of justification not to act as they direct, permissibility-conferring reasons function purely to provide said justification. Such reasons contribute to determinations of rational or moral eligibility, that is, not by issuing a requirement but by defusing or answering the deontic charge of those that do.

Joshua Gert's is the most well known such theory. Though he endorses an analogous theory for morality, his primary interest is in practical rationality. As he argues, one intuitive way of thinking about the principles of practical rationality is to see them as principles that outline classes of action that are irrational absent adequate justification and classes of considerations that can serve *as* that justification. For instance, according to one plausible substantive principle of practical rationality, it is irrational to incur serious risk to self absent sufficient compensating benefits to oneself or others. The adventurousness of sky diving, say, can justify the increment of risk to limb which would be irrational absent any meaningful benefit. To say that the adventure can rationalize the risk of skydiving does not mean it is rationally impermissible *not* to skydive. Instead, the role that consideration plays is to justify the cost rather than requiring the benefit. The considerations that provide justification, that is, can be different from those that place in need of justification.

More technically, for Gert, the foundational notion of practical rationality is the objective, wholesale rational status of action as rationally permissible or impermissible. His central claim is that there are two determinants of such status, which correspond to two normative functions: a "requiring" function, which works toward establishing the rational impermissibility of what would otherwise be rationally permissible, and a "justifying" function, which works toward establishing the rational permissibility of what would otherwise be rationally impermissible.[1] The strength of reasons' requiring force is measured by the scope of actions it could succeed in moving from permissible to impermissible; the strength of reasons' justifying force is measured by the scope of actions it could succeed in moving from impermissible to permissible. A consideration

[1] While this is the definition Gert gives, note that competing requirers would also count as justifiers under this definition. After all, the requiring force of a reason to phi also pushes against competing requirements not to phi; if strong enough, the former will render phi-ing permissible where it would not otherwise have been, as Raz (1975) and others have pointed out, "permissible," after all, is technically consistent being permissible *and* required. But other things Gert says indicates that he does not mean to count competing requirers as playing a justifying role. The analogy to consent helps to clarify that,

may be low in strength as a requirer and high in strength as a justifier, or vice versa. And some reasons are purely justifying – the *only* normative role they play is helping to secure the rational permissibility of certain actions.

Patricia Greenspan independently advances a similar view. Rather than "requiring" and "justifying," she uses the locution of "negative" and "positive" reasons. For Greenspan, the foundational grounding notion of practical rationality (or again morality) is whether or not an action is subject to the grounds of *significant criticism* – that is, forms of criticism, such as "irrational," that rational agents are constitutionally committed to avoiding.[2] Negative reasons are those that subject us to criticism if we fail to do as they direct: that it would be time consuming and irritating, for instance, is a negative reason against joining a faculty committee (amen). Such reasons push to "disqualify" the option in question, namely serving on the committee; in the absence of countervailing considerations, joining the committee would irrational. Positive reasons are considerations that dissipate criticism without introducing any such grounds of their own. That the committee would be helpful to improving one's standing at the university, for instance, is a positive reason. The consideration can rationalize the decision to serve, removing the grounds for criticism it would otherwise merit. But unless the positive reason contains a hidden negative one (one actually needs improved standing to avoid harm and could thus be criticized for passing up the opportunity), such reasons do not open one to criticism if one decides nonetheless not to serve. One *might* be motivated by the potential benefit and decide to join the committee, but one is not (thank goodness) rationally compelled to, for the normative function of the positive reason is simply to remove potential grounds of criticism for the action it concerns.[3] For Greenspan, then, the normative forces of reasons are found in their capability to serve as grounds for offering or for answering criticism of a given option; purely positive reasons provide the latter without constituting the former.

Douglas Portmore endorses and applies Gert's theory in the moral realm. Picking up on comments by Gert, Portmore argues that the requiring/justifying distinction provides a coherent explanation of heroic supererogatory actions, such as running into

---

in fact, Gert means to reserve "justifying" for the function of working to remove a ground of prohibition without introducing one of its own. A better definition of the justifying function for Gert, then, is that it works to render *merely* permissible an act that would otherwise be impermissible. The broader sense of justification, inclusive of competing requirements, is what below we call "defensive justification"; see section 5.

[2] Greenspan explicitly declines to unpack "serious criticism." That said, it is worth pointing out that she endorses the central equivalence of her functions to Gert's. As his functions are clearly about the deontic status of actions, her "serious" criticisms may be ones that track deontic judgments.

[3] Of course, stronger negative reasons can also serve as rejoinders to criticism. Like Gert, Greenspan means to reserve "positive reasons" for reasons that function to dissipate criticism without introducing grounds for criticism. "Negative reasons have the force of O~A [rationally or morally obligated not to A], ... in normative terms, a purely positive reason merely denies this. A positive reason, however strong, at most serves to block a negative reason from binding, unless it really conceals or implies a competing negative reason that is strong enough to defeat it" (p. 363).

a burning building to save another. Intuitively, we are both morally and rationally allowed to sacrifice in these instances while being both morally and rationally permitted not to do so. Rather than defending this intuition by positing, say, normative insulation between moral and personal reasons, Portmore argues that it plausibly follows from reasonable assumptions about when benefits and burdens to ourselves and to others issue requiring force, and when they issue justifying force, in the moral and rational realms. Intuitively, morality faces limits on how much it can demand you to sacrifice; in particular, its requirements should be sensitive to the importance of not demanding an equivalent sacrifice of your own utility, impartially construed. Costs to you thus carry morally justificatory weight to decline altruistic actions that would otherwise be morally required of you.[4] At the same time, those costs to self do not carry moral requiring force to avoid them, for morality, intuitively, is more centrally about requirements to help and avoid harm to others rather than oneself. Relevantly, one does face a pro tanto prudential requirement to avoid the sacrifice – it would be irrational to sacrifice one's life for no adequate compensating good. But here, the benefit that would accrue to others carries clear rational justifying strength: it renders it rationally permissible for you to sustain a loss that would otherwise be irrational. Supererogation is thus both morally and rationally coherent, for it is morally and rationally permissible to make the sacrifice without it being morally or rationally required.

In all of these views, there are two distinct normative roles, not just one, that reasons play in determining the fundamental deontic status of an action – in determining, that is, whether it is irrational (Gert), criticizable (Greenspan), or immoral (Portmore). Requiring or negative reasons function to push certain options – namely, *not* doing as they say – towards the outlaw status in question. Justifying or positive reasons, on the other hand, work to remove that status – by a degree determined by its justifying or positive strength – but without substituting any counter-requirement. In contrast to stronger competing requirements, which render their acts permissible only because now required, justifying or negative reasons serve simply to clear away what would otherwise be a prohibition against it. The permissibility-conferring function, as we might put it, is deontically subtractive.

Core to all of these views is the idea that value does not always ground requiring force. Only selected actions are pro tanto irrational or immoral to do. Practical rationality, for instance, is more about avoiding burdens and losses than about maximizing personal benefits; morality is more about helping others than about helping oneself. Further, and critically, when value does not ground a requirement, it is not normatively inert. Instead, it functions to ground justification, or potential rejoinder, to the

[4] Gert and Portmore endorse weightings that are quite robustly context independent, while Greenspan is skeptical of attempts to formalize the weights of reasons, whether positive or negative, across contexts. Gert thinks that the weighted nature of reasons is a critical aspect of why the concept of "justifying" reasons offers a better account of latitude than either exclusionary permissions or incommensurability, since the latter two, as originally defended by Raz, are not graduated concepts (see Gert, p. 108).

deontic impress of those requirements. One way value does this, as we have seen, is by outlining *compensating benefits* – benefits that are not required but that rationalize a cost and make a tradeoff reasonable. Another way value does this, as we have seen, is by outlining *burdens that set limits* on what it is reasonable to ask an agent to sacrifice on behalf of potential requirers. In both cases, reasons can provide precisely the sort of countervailing considerations the requirement in question is meant to be sensitive to without themselves issuing any deontic push.

As Gert points out, the permissibility-conferring role is thus functionally like consent. Absent your consent, it is impermissible for another to access your body or property; your valid consent removes that prohibition. It does not issue a requirement *that* the person engage in the activity; rather, it simply removes a deontic constraint against doing so. Of course, consent is an authority-based concept: it involves a person deciding to grant permission. Here, the idea is that certain value-based considerations involving burdens, benefits, and the like can impersonally achieve what the activity of validly consenting does personally. They serve to defuse a certain amount of deontic push without issuing any of their own. Such considerations can constitute reasons for action, in a way that the fact of consent does not, because they are also considerations (getting benefits and avoiding burdens) that can motivate creatures like us, while the mere fact that someone gives one permission to phi is, in the usual course of events, not something that itself moves us to phi. But again, in their normative work, such considerations function simply to remove a deontic impediment to the action they concern.

As such, purely justifying (or positive) reasons to phi are not governed by the principle of motivational internalism. One who apprehends a purely justifying (or positive) reason to phi can be fully rational and moral and be entirely unmotivated to phi. While one *might* be moved to act to get the benefit (sky diving) or avoid the burden (staying out of the burning building), that is, it is perfectly consistent with being a fully rational or moral agent that one is entirely unmotivated to do so. Such a response is compatible with full pragmatic uptake of the normative force of the reason, for the normative force of the reason is merely defensive.

4 Commendatory reasons

A second group challenges the functional monism of the standard view in a very different way. Members of this group posit the existence of what we will call a purely commendatory function. Unlike permissibility-conferring, which serves simply, if importantly, to clear away an action's deontic disqualification, commending provides a normative ground *for* choosing an action, just one that carries no deontic charge. While requiring reasons place one in need of justification not to do as they bid, and permissibility-conferring reasons provide that justification, commendatory reasons normatively speak in favor of an action without placing one in need of justification to decline in the first place.

Dancy is perhaps the most well known member of the group. Dancy argues that there are two modes or styles of favoring, which he calls "peremptory" and "enticing."[1] Peremptory favoring is deontic: one stands in need of justification to decline acting as it directs. Enticing favoring, in contrast, "makes its option worth doing," but "what there is overall enticing reason to do will not amount to a reason that it is wrong to act in breach of" (95–6). For instance, that something would be fun to do, he argues, provides normative support for pursuing it, but it would not be irrational, even in the absence of any competing reasons, to decline. Dancy is clear that this is quite different from the function of rendering an action permissible. As Dancy puts it, all that sort of a function does is to establish "the absence of reason against" an action (106). In contrast, enticing reasons are normative grounds for choosing an action, but normative grounds that do not carry with them the grounds of irrationality should one not follow their lead.

For Dancy, this second mode of favoring provides a more straightforward explanation than Raz's appeal to widespread incommensurability for cases of everyday latitude, such as the optionality of foregoing an edifying night at the opera in favor of fun. For Raz, as we noted, all reasons carry a deontic charge. Latitude in such cases must be achieved by the courtesy of finding an opposing but incommensurable requiring reason, which extinguishes the original reason's deontic threat without introducing its own dominant command. Optionality is achieved only by first successfully meeting the deontic charge that every favoring reason brings with it. In contrast, Dancy argues that the explanation of many such cases is more simple: many reasons for action favor

[1] Dancy eschews the term "requiring" in favor of "peremptory" because he believes, for somewhat idiosyncratic reasons, that requiring applies only at the all-things-considered level. He therefore needs a different term to refer to the sort of pro tanto deontic force that applies at the level of individual reasons.

without carrying any deontic charge to begin with. They render their own option pro tanto choiceworthy to pursue, but one doesn't need the excuse of another to decline: one simply has the option, with respect to each, or both, to walk on by.

Horgan and Timmons defend the existence of non-deontic favorers in the moral realm. They focus on the supererogatory nature of small kindnesses, such as offering to take a recently widowed neighbor to a ballgame because you know it would mean so much to her. Exploring the phenomenology of encountering such a reason, they note that it does not bring with it any felt need to offer an excuse not to issue the invitation, and brings with it "no sense that guilt, shame, or blame would be appropriate" should one decline – the phenomenological marks of requiring force (48–49). That said, the experience does bring with it a sense of helping as *worth* doing, as a *meritorious* end – the phenomenological marks of normative favoring. Having argued for the importance of taking phenomenology seriously, they argue that it mirrors a distinction in the types of favoring that reasons can bring. Some reasons provide normative grounds that favor in a requiring mode, but some favor without issuing any, even pro tanto, requiring force.

Like Dancy, Horgan and Timmons are explicit that this function is not the same as "justifying" in Gert's sense of the term. In the example at hand, the moral reason to issue the invitation is not functioning to help secure the moral permissibility of doing so – after all, there was no competing requirement against doing so. Instead, such reasons serve to provide normative grounds for pursuing an action but grounds that build in the optionality of declining. Citing the goods that can come from acting in worthy ways that are not required, these "moral merit-conferring reasons" speak in favor of an action but with a normative force that is evaluative rather than deontic.

Fiona Woollard (2016) also defends the existence of non-deontic reasons in the moral realm. Her central concern is with a problematic social reasoning around maternal obligations, namely the tacit view that mothers must justify declining each instance of benefiting their children. In challenging the view, she argues that one assumption that often lies behind it is the idea that a moral reason to do something must imply a "defeasible duty" to do so. Citing and extending Dancy's concept of enticing reasons, she argues that moral reasons need not always place one in a position of needing justification not to act in their service. To think so is to get wrong the structure of the reasons. They endorse and provide a normatively worthy basis – indeed, a moral basis – on which one can act, but one does not need to gather justification or assemble permission not to take each instance. One does not need to earn one's way out of every opportunity to advance the interests of one's children, even as those interests provide a moral reason that can serve as a normative basis to which one can respond.

We also have defended the existence of a distinct "commendatory force," which favors in a non-deontic style, and offered a framework for thinking about the normative interaction effects of such reasons with one another (Little and Macnamara 2017).

In all of these views, there are two, not just one, modes in which reasons for action can serve to normatively favor or endorse an action. Requiring (peremptory, obligating)

force does so by making it pro tanto wrong not to perform the action. In the absence of countervailing factors, the action is endorsed by the very good fact that it would have been wrong not to do. Commendatory (enticing, merit-conferring, evaluative) force does so by making the action pro tanto *worth* doing. Commendatory reasons serve to expand the set of worthwhile options without expanding what is required of us. Commendatory reasons favor actions in a way that internally maintains the optionality of declining. If requiring is a normative stick, commending is a normative carrot.

As we've argued elsewhere (Little 2013), if requiring functions like a demand, and permissibility-conferring like consent, commending thus functions structurally like a request. Like a demand, to issue a request is to exercise a form of standing to interpersonally insert oneself into the other's normative considerations: where felicitous, such speech acts provide one a normative, second-personally sourced ground for doing the act in question (see Enoch 2011, Darwall 2013). That said, requests have a fundamentally different structure than demands. While demands place a requirement that their recipient act as directed, absent specific justification, requests do not presume to give their recipients an assignment, even a pro tanto one. In the standard case, one needs to provide a response, but 'yes' and 'no' are both fully acceptable answers (Raz 1999). While other things may contingently make it wrong to decline, the request itself does not place one in need of exculpatory justification to decline, even as it gives one a normatively adequate basis for acceding.

Core to the defense of commendatory reasons is the claim that certain value-based considerations can impersonally achieve what the activity of validly requesting does personally. The good we can do is not always our job to attain, placing us in need of exculpation if we decline to act in its service, but the choiceworthiness remains.[2] Commendatory reasons set out the worthiness of the end but with built in optionality. As such, purely commendatory reasons to phi are not governed by the principle of motivational internalism. However strong a commendatory reason, so long as it is merely commendatory, a fully rational agent can gaze upon it in full appreciation and simply stay on the sidelines, declining its request. Reasons can favor actions without any requiring force at all, for there are evaluative, not just deontic, modes of favoring.[3]

[2] Dancy himself believes that mild (non-deontic) criticism is apt when one chooses to follow a weaker over a stronger enticing reason (pp. 92–93, 103–104). For a contrasting view, see Little and Macnamara (2017), where we provide an account of comparative commendatory dominance that implies no criticism at all in choosing the lesser.

[3] As Horgan and Timmons say, "to identify the kind of role one needs, one must look to non-deontic forms of moral evaluation. It will not work to focus only on deontic evaluation and the roles that reasons play as they bear on the deontic status of actions" (pp. 53–54).

5 Conclusion

The previous has outlined two very different claims about non-requiring functions that practical reasons may play. The first serves to render permissible; the second serves to commend.

Yet some in the literature have struggled to keep clear the difference between these functions. To give just one example, in his defense of "optional reasons," Scanlon (2014) registers his concurrence with Gert, but his own discussion looks far more like a defense of commending than anything about permitting.

Why the confusion? One reason, we suspect, has to do with the term "justifying." As we've seen, Gert (and Portmore following him) uses the term "justifying" to describe the non-requiring function he endorses. But justification is a concept that contains an inherent ambiguity. When Gert and Portmore use the term, they mean justifying in what we might call the defensive or exculpatory use of the term – defending against or disabling a negative charge that would otherwise obtain. This is indeed an important concept of justification. To justify an action in this sense is to move the action off the "No Fly List" of rationally or morally forbidden actions. When Gert and company say that reasons can justify without any requiring force, they mean that reasons can recover the permissibility of an action by means other than issuing a stronger competing demand.

But justification is also widely used in a very different sense, to refer to establishing an action as something that pro tanto ought *to* be done, as pro tanto *worthy* in some respect, or participates in a form of *good*. To say an action is justified in this sense is to say more than that it is not forbidden; it is to say that pursuit of it can be grounded in more than inclination; it is endorsed by something that normatively speaks in favor of doing it. This is the sense of justification – what we might call justification in the endorsing sense – that commendatory theorists are interested in. When they say that reasons can justify without issuing requiring force, they mean that reasons can normatively endorse an action in a non-deontic manner.

To be clear, then, while both sides claim, for instance, that goods or benefits can "justify" more than they require, they mean very different things by "justify."[1] The first

[1] The distinction between the two senses of justification can also get obscured because Gert categorizes his "purely justifying" reasons – that is, reasons whose only normative function is to help render an action permissible – as *favoring* that action. This, it should be said, is an idiosyncratic use of the term "favorer." When most people use the term – certainly when Dancy or Horgan and Timmons or we use the term – they precisely mean a consideration that justifies an action in the normative endorsement sense of the term. For Gert, in contrast, it is enough for a consideration to count as a favorer that

camp means that goods or benefits can provide more defensive justification for enduring a risk or burden than they ground requirements to attain the goods themselves. The second camp means that goods or benefits can commend a broader scope of actions than they require. The first claim remains staunchly within the deontic life of reasons – the realm of what is permissible versus forbidden. It works to expand the ways in which reasons contribute to determining which of those statuses an action has, adding a deontically defusing function to the familiar requiring one. The second claim leaves the deontic realm altogether. It is interested in the normatively endorsing role of reasons – the ways in which reasons underwrite a normative grounds for choosing an action. Their interest is in expanding the ways in which reasons can favor an action as one worthy of choice, from one that is deontic to one that is evaluative.

Of course, underscoring the difference between the two non-requiring functions does not mean one can't believe in both. It is also possible to belong to *both* camps, and many do. Horgan and Timmons, for instance, explicitly endorse the legitimacy of both non-requiring functions. While Portmore concentrates his efforts on defending the existence of a permissibility-conferring function, he also declares himself open to a commendatory function. Indeed, one can believe that one and the same consideration can carry all three forces – requiring, permitting, and commending. For instance, one view of heroic supererogation (as opposed to its quotidian cousin) is to say precisely that. That one could save another's life carries morally deontic force; absence sufficient countervailing justification, one would be wrong not to act. The risk to one's life provides said justification, and declining is permissible. That said, the good that can come of saving the life serves not only to justify, in the Gertian sense, what would otherwise be an irrational danger to one's well-being, it is very much in view as a morally commendatory reason *to* do so.

Still and all, it would be a mistake to say that everyone is bipartisan. For instance, it is easy to believe in the permissibility-conferring function of reasons and not believe in the commendatory function. In fact, two of the three theorists we pointed to as

---

it be something that can in fact motivate creatures like us, so long as it plays *some* role – even a role in removing a deontic impediment – in adjudicating the objective status of actions as impermissible. Such a consideration might better be thought of less as a traditional favorer and more of as a motive that also happens to serve as a truth condition of permissibility.

Greenspan's locution of "positive reason" to refer to a permissibility-conferring reason carries similar potential to mislead. The two senses of justification outlined in this section are sometimes referred to as negative and positive justification. Her invocation of "positive reasons" can carry a connotation of positive justification – that is, of a normative basis to take the action it concerns, whereas, as she herself emphasizes, "positive reasons" simply signal removal of a reason against doing its action.

Disambiguating the two senses of justification can help give technical definitions of the three normative forces. The function of requiring is to provide endorsing justification by way of placing one in need of defensive justification not to do as it acts. The function of permissibility-conferring is to provide defensive justification against a competing requirement and does not itself provide any endorsing justification for the action it concerns. The function of commending is to provide a kind of endorsing justification that does not place one in need of defensive justification.

endorsers of permissibility-conferring, Gert and Greenspan, are skeptics about commendatory force. They are what we might call deontic reductivists about normative support. There is nothing normative to the concept of 'merit' or 'worthy' other than grounds for requiring or for permitting. If one is moved to act on the basis of a purely permitting reason, it is a matter of plain inclination, not anything normative. (As Greenspan puts it, the "pull" of purely positive or justifying reasons is only motivational, not normative [390–91].) They see the normative economy of reasons, that is, as thoroughly and exhaustively deontic.

In contrast, the core claim of those defending purely commendatory reasons is that there is more to the normative realm than the deontic. The normativity of practical reasons can speak in favor of an action beyond either requiring it or clearing away a prohibition against it. For the normative is not solely about the deontic, it is also about what is good, merited, or choiceworthy. The normative lives of reasons are not just about imposing or answering deontic vulnerability, for the deontic does not exhaust the normative.

Those who defend commendatory force are less likely to be skeptics writ large about the ability of non-requiring reasons to play a permissibility-conferring function. After all, if one believes a consideration renders an action worthy to pursue, it is likely one will believe it capable of rationalizing at least some risks or burdens, at least in assessments of garden-variety rationality. Even if a benefit is not deontically required, that is, if one thinks it renders an end worthy it would be odd not to think that worthiness could justify some risk or burden.

That said, it is important here, too, to keep the distinction between the functions clear, for they are unlikely to be fully co-extensive. Certainly, the fact that a consideration acts as a commendatory reason to phi in some circumstance does not mean it is serving there to help render it permissible to phi. As we saw with Horgan and Timmons' example, to say the reason to issue the invitation to the ballgame is commendatory does not mean it is also helping to render the invitation morally permissible, for there may not be any threatened counter-requirement standing in way of the invitation (say, a standing obligation for your time) to begin with. More generally, one might include considerations in one's list of commendatory goods without regarding them as the kind of things that could push back on one's favored list of pro tanto rational prohibitions. One might believe that fun is an important good capable of conferring merit and commending action but not regard it as the right sort of good to rationalize, say, a risk to life and limb. In morality, too: one can certainly think something a moral good yet not think it capable of getting one out of a given moral duty.[2]

But the central point, at any rate, is that these are distinct functions and distinct claims. Both are important challenges to the functional monism of the standard view

[2] This, in essence, is the topic of Frances Kamm's famous 1985 article on supererogation. One can believe in non-required moral goods' commendatory force but believe they are never able to push back against pro tanto moral requirements. Kamm's article defends the opposite.

of reasons. Debates between members of the camps can be as important as debates between any camp and the standard view – and each deserves its own discussion.

References

Audi, R. (1997). *Moral Knowledge and Ethical Character.* New York: Oxford University Press.

Chang, R. (1997). Introduction. In R. Chang (Ed.), *Incommensurability, Incomparability, and Practical Reason* (pp. 1–34). Cambridge, MA: Harvard University Press.

Copp, D. (1995). *Morality, Normativity, and Society.* New York: Oxford University Press.

Dancy, J. (2004). Enticing Reasons. In R. J. Wallace, P. Pettit, S. Scheffler, and M. Smith (Eds.), *Reason and Value: Themes from the Moral Philosophy of Joseph Raz* (pp. 91–118). Oxford: Clarendon Press.

Darwall, S. (1983). *Impartial Reason.* Ithaca: Cornell University Press.

———. (2013). Morality and Principle. In D. Bakhurst, B. Hooker, and M. Little (Eds.), *Thinking About Reasons: Essays in Honor of Jonathan Dancy* (pp. 168–191). Oxford: Clarendon Press.

Edgley, R. (1965). Practical Reasoning. *Mind* 74, 174–91.

Enoch, D. (2011). Giving Practical Reasons. *The Philosopher's Imprint* 11(4).

Gert, J. (2004). *Brute Rationality.* Cambridge: Cambridge University Press.

Greenspan, P. (2005). Asymmetrical Reasons. In M. E. Reicher and J. C. Marek (Eds.), *Experience and Analysis: Proceedings of the 27th International Wittgenstein Symposium* (Vienna: Austrian Wittgenstein Society), pp. 387–94).

Horgan, T. and Timmons, M. (2010). Untying a Knot from the Inside Out: Reflections on the "Paradox" of Supererogation. *Social Philosophy and Policy* 27(3), 29–63.

Kagan, S. (1989). *The Limits of Morality.* Oxford: Clarendon Press.

Kamm, F. (1985). Supererogation and Obligation. *Journal of Philosophy* 82(3), 118–138.

Kauppinen, A. (2015). Favoring. *Philosophical Studies* 172(7), 1953–1971.

Kolodny, N. (2003). Love as Valuing a Relationship. *Philosophical Review* 112 (2), 135–189.

Korsgaard, C. (1996). *The Sources of Normativity.* Cambridge: Cambridge University Press.

Little, M. (2013). In Defense of Commendatory Reasons. In D. Bakhurst, B. Hooker, and M. Little (Eds.), *Thinking About Reasons: Essays in Honor of Jonathan Dancy* (pp. 112–136). Oxford: Clarendon Press.

Little, M. and Macnamara, C. (2017). For Better or Worse: Commendatory Reasons & Latitude. In M. Timmons (Ed.), *Oxford Studies in Normative Ethics*, Vol. 7 (pp. 138–170). Oxford: Oxford University Press.

Portmore, D. (2012). Commonsense Consequentialism: Wherein Morality Meets Rationality*. New York: Oxford University Press.*

Raz, J. (1975). Permissions and Supererogation. *American Philosophical Quarterly* 12(2), 161–168.

———. (1989). *The Morality of Freedom*. Oxford: Oxford University Press.

———. (1999). *Practical Reason and Norms*. Oxford: Oxford University Press.

——*. (2000).* Engaging Reason: On the Theory of Value and Action*. Oxford: Oxford University Press.*

Scanlon, T. M. (1998). *What We Owe to Each Other*. Cambridge: Harvard University Press.

———. (2014). *Being Realistic About Reasons*. Oxford: Oxford University Press.

Smith, M. (1994). *The Moral Problem*. Oxford: Wiley-Blackwell.

Tannenbaum, S. (2007). Brute Requirements: A Critical Notice of Joshua Gert's *Brute Rationality*. *Canadian Journal of Philosophy* 37, 153–172.

Velleman, D. (1996). The Possibility of Practical Reason. *Ethics* 106, 694–726.

Woollard, F. (2016). Motherhood and Mistakes About Defeasible Duties to Benefit. *Philosophy and Phenomenological Research* 97, 126–149.

27. Requirements of reason

R. Jay Wallace

It is natural to think that some normative considerations represent requirements of reason. They do not merely speak in favor of actions or attitudes but demand or require that we adopt them. It is a somewhat neglected task for the theory of practical reason to make sense of the standing of at least some normative considerations, including those at the center of morality, as requirements in this sense. I shall canvas several approaches to this problem that are implicit in the philosophical tradition and test their strengths and limitations, paying particular attention to the controversial case of moral obligation. The discussion will consider identity-based conceptions of a practical requirement, dominant reason models, perfectionist accounts, and approaches that are grounded in the relational character of certain reasons. I shall suggest that the relational model is especially promising when it comes to the case of moral obligation but also that we should acknowledge a plurality of sources of practical requirements.

1 Requirements of reason: the basic idea

Many normative considerations exhibit what I have elsewhere called *deontic structure* – that is to say, they constitute requirements to do some things and not do others.[1] It is common to understand morality in these terms: it does not merely *recommend* certain actions but rather presents us with obligations which constrain our thinking about what to do. While this chapter focuses mostly on the moral case, it is worth noting that some other normative domains have been understood in similar terms. To take an example discussed extensively elsewhere in this handbook (see, e.g., Lord's contribution), *rationality* has often been understood as a source of requirements which prohibit or demand certain combinations of attitudes rather than merely recommending certain attitudes as eligible options.

Practical requirements or obligations – I will use these terms interchangeably in what follows – have an importantly different normative status from other reasons, most centrally including what Jonathan Dancy has called *enticing reasons*.[2] To take a standard example, consider the reasons given by considerations of pleasure. The fact that it would be fun for me to learn how to dance is a reason to pay for dancing lessons. But even if this reason seems compelling to me and it is not outweighed, I will not regard it as *obliging* me to pay for dancing lessons. By contrast, if I promised that I would pay for your dancing lessons, it will not seem similarly optional to do so: I will feel that I have acquired an obligation to do so. In this way, practical requirements have what Samuel Scheffler has called a "presumptive decisiveness".[3]

Attempts to characterize normativity as a simple function of the *balance of reasons* sometimes overlook these distinctive features of practical requirements. It has become common to explain normativity entirely in terms of pros and cons that weigh against each other to yield overall verdicts about what to do.[4] Admittedly, reasons often work

[1] See my paper "The Deontic Structure of Morality", in David Bakhurst, Brad Hooker, and Margaret Olivia Little, eds., *Thinking About Reasons: Essays in Honour of Jonathan Dancy* (Oxford: Oxford University Press, 2013), pp. 137–167.

[2] See his *Ethics Without Principles* (Oxford: Oxford University Press, 2004). For a fuller discussion of these reasons, see the contribution to this volume by Little and Macnamara.

[3] See Scheffler, "Relationships and Responsibilities", as reprinted in his *Boundaries and Allegiances: Problems of Justice and Responsibility in Liberal Thought* (New York: Oxford University Press, 2002), pp. 97–110, at p. 100.

[4] See T. M. Scanlon, *What We Owe to Each Other* (Cambridge, MA: Harvard University Press,

in just this way: to continue the earlier example, the fun of learning how to dance counts in favor of taking dancing lessons, but the time and money consumed by these lessons may tip the scales in favor of a different pastime. But practical requirements work in an importantly different way. On the face of it, they are not put on the scales alongside the attractive and unattractive features of an action but rather serve as *exclusionary reasons* (to use Joseph Raz's influential phrase): they can exclude certain reasons that would otherwise be in the running.[5] In other words, requirements function to preclude other reasons from being placed on the scales in the first place as considerations that are to be weighed against each other.

Another difference between practical requirements and considerations that "count in favor" of actions is that we have a degree of latitude in responding to the latter but not the former. It is often a matter of our discretion whether to pay heed to the balance of considerations that merely favor some option. There may, for example, be no option which stands out as best. Different kinds of reasons might provide good cases for several different options, and an agent might select one while acknowledging that there are considerations that speak in favor of the others. Even if there is a best option, there may still be other options for which good enough reasons can be given, and we may have discretion to choose among these good enough options. Cases governed by practical requirements stand in stark contrast: nothing short of compliance would qualify as good enough, and we have no latitude to discount or ignore them in deliberation. Partly for this reason, we express such requirements in ordinary talk with peremptory language like "must" and "have to", rather than with a mere "ought".

Owing to these differences, requirements of reason enter into our practical reflections in a distinctive guise. The natural way to register them in deliberation is not to weigh them against normative considerations of other kinds but to form future-directed intentions to comply with them. As many have recognized in the literature,[6] these intentions constrain our practical deliberation by *resolving* the question of whether to do X, so that future deliberation proceeds in light of the assumption that X will be done. The practical requirement to X is not taken into account as just one consideration among others. Rather it is already a sufficient basis for forming the intention to do X, which narrows the practical horizon by removing from the agenda options inconsistent with one's future X-ing.

To be sure, intentions that structure practical reasoning in this way are not formed only in response to practical requirements. For example, one might settle on using one's

---

1999), Chapter One, for the idea of reasons as considerations that count in favor of attitudes; and Joseph Raz, "Explaining Normativity: On Rationality and the Justification of Reason", in his *Engaging Reason: On the Theory of Value and Action* (Oxford: Oxford University Press, 1999), pp. 67–89, at p. 67, for the idea that normativity is to be understood fundamentally in terms of reasons.

[5] See Joseph Raz, *Practical Reason and Norms*, Second Edition (Princeton: Princeton University Press, 1990).

[6] The classic statement of this view is Michael Bratman, *Intention, Plans, and Practical Reason* (Cambridge, MA: Harvard University Press, 1987).

quiet evenings in the next few months to learn about the Maurya Empire rather than reading the novels of neglected Modernist writers. This decision may then exclude one's reading Hermann Broch on Tuesday evening but not because an antecedent requirement is in place to study South Asia during the Iron Age. The point is rather that when there is a practical requirement to do X, it is a sufficient basis on its own for forming a future-directed intention to do X, even if one hasn't engaged in any weighing of the pros and cons that bear on X-ing. Suppose, for example, that I had instead promised that I would participate in your book club, and the club is spending the next several months on underappreciated Austrian novels from the first half of the 20th century. Then I may be obliged to read Broch and postpone learning about ancient India even if I haven't compared these options and would have found the latter more enjoyable had I done so.

Although practical requirements do function to constrain one's practical horizon, it is worth emphasizing that they are not *absolute* but represent only *presumptive* constraints. Their force can be defeated under certain conditions, and in this respect their normative profile overlaps with that of more ordinary reasons. For example, there may be circumstances that *disable* a practical requirement, such as the unanticipated emergency that cancels an obligation incurred by a promise. Here the requirement ceases to apply. In other cases, the requirement may remain in force, but one might find oneself in the unfortunate situation that one cannot comply with it without violating another requirement that is in force. Moral dilemmas, if there are any, may provide examples of this kind of case. Such cases are also discussed in the literature on rational requirements. If, for example, we should adopt the *narrow scope* view of coherence requirements discussed in Lord's contribution to this volume, it may be possible for rationality to paint one into a corner in which one's options all involve some form of irrationality. Another example within this domain would be a conflict between the attitudes supported by one's higher-order evidence and the attitudes supported by one's first-order evidence. In such cases, it could be defensible in retrospect to ignore the higher-order evidence if it becomes clear that it was misleading, even if it would have been irrational at the time to do so.[7] In any of these cases, one may need to face up to the fact that there is no path forward in which one does not do something forbidden.

Future-directed intentions are well suited to operationalize the recognition of requirements as presumptive constraints. These intentions resolve the question of whether to do X in a way that constrains deliberation, but they do this with a flexibility that matches the defeasibility of presumptive constraints. While it is part of their function that future-directed intentions tend to resist reconsideration, circumstances can arise in which it is rational to revise them, as the literature on intention has long recog-

[7] For a discussion of such apparent dilemmas, see Errol Lord and Kurt Sylvan's, "Suspension, Higher-Order Evidence, and Defeat", in J. Brown and M. Simion, eds., *Reasons, Justification, and Defeat* (Oxford: Oxford University Press, forthcoming).

nized.[8] For this reason, plans are typically formed with a built-in sensitivity to ways in which new developments can give us reasons to reconsider them. This sensitivity of rational intentions to the possibility of reasons for reconsidering mirrors the way in which requirements function to define defeasible constraints on agency, determining what is to be done in circumstances that might well change going forward, so as to undermine their normative force.

[8] See, again, Bratman, *Intention, Plans, and Practical Reason*. See also Bratman's contribution to this volume.

2 Explaining requirements of reason

The preceding considerations suggest that practical requirements represent normative relations that are worth distinguishing from the favoring relations that hold between recommending reasons and actions. The apparently *sui generis* character of practical requirements raises an important philosophical question: what features of obligations render them suited to support conclusions that are expressed using the deontic "must" and to structure deliberation in the way of presumptive constraints? There are three familiar strategies for answering this question in the philosophical tradition, which I will call the *dominance model*, the *identity-based model*, and the *voluntarist model*.

The dominance model seeks to explain the stringency of requirements in terms of reasons that are robust – that is, sufficient, in a wide range of practical contexts, to outweigh reasons on the other side of the ledger. In this model, requirements are not different *in kind* from other reasons. They just have more weight in more contexts so that they routinely *dominate* other reasons. It is the fact that the normative grounds of requirements are robustly weighty in this way that justifies us using the deontic "must", and that explains our tendency to treat them as presumptive constraints. This way of grounding requirements in robust reasons has a long legacy in the literature on the rationality of morality, but it has also come to the fore in some recent work on requirements of rationality: Benjamin Kiesewetter and Errol Lord, for example, have argued that requirements of coherence – which demand certain combinations and ban other combinations of attitudes – can be explained by appealing to the fact that a person's total set of reasons will almost always favor attitudes that cohere over attitudes that do not (though the particular attitudes in these coherent sets will be favored by different reasons in different cases).[1]

The identity-based model, by contrast, seeks to explain requirements by means of normative considerations that trace to our most significant self-conceptions. This model is most familiar from the work of Christine M. Korsgaard, who has applied it to explain both moral requirements and rational requirements.[2] To see an example of this model in operation, consider one's self-conception as a parent. If you are, and

[1] See Benjamin Kiesewetter, *The Normativity of Rationality* (Oxford: Oxford University Press, 2017); Errol Lord, *The Importance of Being Rational* (Oxford: Oxford University Press, 2018).

[2] Korsgaard develops this approach in several works: *The Sources of Normativity* (Cambridge, MA: Harvard University Press, 1996); *Self-Constitution* (Oxford: Oxford University Press, 2009), and the essays in *The Constitution of Agency* (Oxford: Oxford University Press, 2008). Other versions of this approach are developed in Bernard Williams, *Ethics and the Limits of Philosophy* (Cambridge, MA: Har-

understand yourself to be, someone's mother – and if this identity is bound up with your sense of the meaning of your own life – then you will naturally experience the needs of your child as making practical demands on you and respond to them in that way in making decisions about how you go about your life. A failure to do this would represent an existential threat to your conception of who you are. As this example illustrates, practical identities lead one to treat certain considerations as having special force, insofar as compliance with them is necessary to sustain the identities with which they are connected. The basic idea behind the identity-based model is that at least some practical requirements are constitutively connected in this way to our most fundamental conceptions of ourselves.

The third approach – voluntarism – seeks to ground requirements in relations of authority, such as the commands of a suitably constituted legislator. A familiar example of this approach is the divine command theory, which treats the moral requirements that apply to us as commands of a supremely authoritative divinity. But there are more mundane relations of authority that could be invoked to provide a voluntarist account of requirements. The relation of subjects to the sovereign in Hobbes's *Leviathan*, the relation of citizens to duly constituted legislative assemblies, professional relations, and familial relations are all contexts in which individuals stand in authority relations to each other of the kind that can give rise to requirements or obligations. Voluntarism provides a transparent account of the distinctive deontic character and role of practical requirements in contexts of this kind: the commands of a legitimate authority are naturally treated as presumptive constraints on the conduct of those who are subject to the authority.

These three philosophical approaches to obligation are not mutually exclusive; different practical requirements might have different grounds. Most obviously, one might apply different models to different normative domains: one might, for example, coherently interpret requirements of rationality in accordance with the dominance model while preferring a voluntarist account of moral requirements. One might also apply different models within a single normative domain; a theorist who accepted a plurality of *prima facie* duties in the style of W. D. Ross might, for example, give different accounts of their grounds which conform to more than one of these models. I will return to this issue in the conclusion; in the meantime, I should like to consider some potential challenges that each model faces, which should help us to understand better their real strengths and weaknesses.

There are at least two significant problems for the dominance model, both of which flow directly from its attempt to understand obligations in terms of reasons that are

---

vard University Press, 1985), chap. 4; in Bernard Williams, "Moral Incapacity", as reprinted in his *Making Sense of Humanity and Other Philosophical Essays 1989–1993* (Cambridge: Cambridge University Press, 1995), pp. 46–55; Harry Frankfurt, *Necessity, Volition, and Love* (Cambridge: Cambridge University Press, 1992). For an illuminating discussion of volitional necessity, see Gary Watson, "Volitional Necessities", as reprinted in his *Agency and Answerability: Selected Essays* (Oxford: Clarendon Press, 2004), pp. 88–122.

robust in a wide range of deliberative contexts. The first problem derives from the scalar character of weight. Given that robustness is a scalar notion, the dominance model entails that the difference between requirements and mere recommending reasons should be a matter of degree. But this seems wrong: requirements seem to us different from ordinary reasons not merely in degree but in kind.

Against this, it might be replied that the notion of dominance is not itself scalar, even if it rests on scalar comparisons in various different practical contexts. But how exactly is the qualitative notion of dominance to be defined? To answer this question, we would need a non-arbitrary account of the threshold above which ordinary reasons turn into practical requirements; but an account of this kind seems elusive. In addition to this problem, the dominance model also seems to rule out dilemmas too quickly. If requirements are understood as considerations that robustly outweigh the other reasons with which they might compete, then it seems that a conflict of requirements cannot arise: by definition, a normative consideration that is comparable in weight with its competitors in a given situation cannot be a requirement, according to the dominance model.

The identity-based view faces different problems. One difficulty involves situations in which we flout requirements while continuing to recognize their force. According to the identity-based view, a requirement applies to me in virtue of the fact that, were I not to comply with it (one might perhaps add "often enough"), some part of my identity will be threatened or even lost. Yet it seems possible to fall short of requirements that we acknowledge to obtain – even to do so with some regularity. It is difficult to understand how this could be possible on the identity-based model: if failing to comply undermines a practical identity of mine, then it will no longer be present to give continued violations any significance. The fact that I am able to flout the obligation will show that it lacks the significance for my self-conception that it would need to generate a requirement.[3]

The voluntarist model avoids these difficulties. By appealing to commands, it invokes a practical consideration that is different in kind, and not merely in degree, from the grounds of recommending reasons. Furthermore, it is clearly possible for a person to be subject to conflicting commands from different authorities (or even the same authority), so the account doesn't rule out the possibility of practical dilemmas. Finally, a person can easily fail to obey a command while nonetheless appreciating its force, which is a template for understanding the phenomenon of flouted obligation. These advantages might account for the enduring appeal of this kind of view in the history of philosophy.

There are, however, other concerns about the voluntarist model. The model applies only in social situations, such as those described in Hobbes's political philosophy, that satisfy the following desiderata. First, there needs to be an individual or group that

[3] For a more general version of this objection to constitutivist views (of which the identity-based view is a special case), see the contribution to this volume by David Enoch.

is recognized to have authority over the subject population. Second, it also needs to be the case that this figure or group has actually exercised their authority by issuing a directive to the subject population. While some practical obligations can no doubt be explained in this way, there are others that cannot be made sense of in these terms, including the obligations at the heart of morality.[4]

[4] One way to try to adapt the voluntarist model to the moral case would be to internalize the relation between sovereign and subject, so that the person issuing commands and the person subject to them are one and the same rational agent. This approach finds expression in Kant's "ethics of autonomy", which conceptualizes the moral law as a principle that rational agents prescribe for themselves whenever they act. There are at least two serious challenges that confront this approach, however: First, there is the general question of whether the authority relation can really be internalized in the way the approach requires (how can a practical law be binding on me if I can always rescind it when I don't feel like complying?). Second, there is the specific challenge of showing that a substantive commitment to the moral law is implicit in the structure of rational agency. For some relevant discussion, see my "Constructivism about Normativity: Some Pitfalls", in James Lenman and Yonatan Shemmer, eds., *Constructivism in Practical Philosophy* (Oxford: Oxford University Press, 2012), pp. 18–39.

3 Explaining moral requirements

The task of explaining the normativity of moral requirements raises particular problems that don't arise for practical requirements as such. These problems have led at least one prominent philosopher, G. E. M. Anscombe, to question the modern notion of moral obligation and to argue that we lack the framework of ancillary ideas that would make it intelligible.[1] To begin with, an account of moral obligation must explain why specifically moral requirements represent considerations that properly enter the deliberative field in the guise of presumptive constraints on agency. There are philosophical controversies about how the class of moral considerations should best be characterized. For ease of exposition in what follows, however, we may provisionally assume that the most significant moral considerations are those that are expressed through the most general expressions of summary moral assessment, namely "right" and "wrong". A non-debunking account of moral rightness, then, must make sense of its apparent deontic structure, the standing of conclusions about what it is morally right and wrong to do as practical requirements of the kind discussed previously.

A second challenge confronts modern conceptions of morality in particular, which typically conceive of the moral community in a maximally inclusive way. In this inclusive approach, all persons are taken to be deserving of equal consideration, as members of the notional community of moral individuals. There are different views that can be taken about whether the moral community extends beyond these boundaries, to encompass (for instance) non-human animals who are capable of pleasure and pain, but the modern approach takes the interests of persons (at a minimum) to count equally in reflection about what it is right or wrong for agents to do. Furthermore, the reason-giving force of conclusions about what is morally moral right and wrong, for moral agents, is partly explained by their membership in an inclusive community of this kind. An account of moral normativity, then, needs to explain the status of these moral considerations not merely as presumptive obligations on the rational will but as obligations that somehow have their source in our common membership in an extensive community of creatures with equal standing.

It is a continuing challenge for moral theory to make sense of these salient features of the moral domain. To illustrate the difficulty, let's consider how three familiar moral theories fare with respect to them, starting with act consequentialism. In its classical form, act consequentialism explains the moral rightness of an action in terms of its maximizing utility, where utility is interpreted in terms of the amount of pleasure and

[1] See her "Modern Moral Philosophy", *Philosophy* 33 (124): 1–19, 1958.

pain produced by the act. It is a significant strength of classical utilitarianism that it does justice to the inclusiveness of morality, taking the pleasures and pains of all sentient creatures to count equally in assessing the impartial value of the states of affairs produced by an agent's actions. But it is much less clear that act consequentialism can explain the character of moral obligations as presumptive requirements.

As we saw earlier, there are different models we can appeal to in order to make sense of the character of normative considerations as practical requirements. But none of the familiar models seems to fit the utilitarian conception of moral rightness. Consider, for instance, the dominance model, which holds that requirements are in place when there are ordinary reasons that robustly dominate their competitors in many different contexts of practical reflection. Utilitarian rightness does not seem to be a source of practical requirements of this kind, however. For an action to be right in act consequentialist terms, all that is required is that it lead to *better* consequences than the available alternatives. But an action could meet this standard even if its consequences are only marginally better than the consequences of the alternatives. Under these circumstances, the balance of reasons in favor of the right action could hardly be said to dominate the reasons on the other side, and so we could not understand utilitarian rightness as a source of obligations according to the dominance model. Nor do the other models of a practical requirement that we canvassed earlier seem to apply any better to the utilitarian conception of moral rightness. We are left in the dark as to why considerations of moral right and wrong, interpreted in utilitarian terms, should have the distinctive force of practical requirements.

The next approach I would like to consider is the divine command theory, which holds that acts are morally wrong if and only if God commands us not to perform them. This approach offers a straightforward account of the obligatory force of moral considerations, insofar as it directly applies to them the venerable voluntarist model of a practical requirement. It is perhaps for this reason that Anscombe considered the divine command theory to be the paradigmatic framework for making sense of the notion of moral obligation. (Her complaint about it was not that it fails to explain the character of moral rightness as a source of obligations but that its theological assumptions are no longer taken for granted in the contemporary world.)

But the divine command theory, even if one accepts its theological underpinnings, seems to do less well with the inclusive aspect of modern morality – the idea that moral requirements derive somehow from our common membership in a notional community of moral equals. God might perhaps command us to treat each other as equals in some sense or other, enjoining that we show respect and consideration for all persons, or that we understand ourselves to be equally subject to the divine will. But these are accidental features of the content of God's commands, not features that are built into the conception of what makes something a moral obligation in the first place. For all the divine command theory says, God might equally have issued commands that have nothing to do in their content with the equality of individual human subjects (such as arbitrary dietary restrictions) or that even deny such equality (favoring one tribal

group over others). In this respect, the divine command theory doesn't really do justice to a central tenet in the characteristically modern conception of moral requirements.

The last familiar theory I would like to consider is a neo-Aristotelian version of perfectionism, which grounds ethics in facts about what Phillipa Foot calls "natural goodness."[2] According to this approach, an action is morally wrong if its performance is incompatible with the traits that human beings need in order to be good *qua* human beings, and morally right if individuals with those traits would perform it. Individuals count as good *qua* human, in turn, insofar as they have the traits that members of their species need in order to flourish or do well under the normal circumstances that they encounter.

The appeal of this approach, when it comes to illuminating the deontic structure of morality, derives in part from the fact that it is a version of the identity-based model. The fact that an individual belongs to a determinate life-form is inescapable, part of what it is to be an individual organism in the first place; hence the standards associated with the life-form apply to individuals who participate in it non-optionally. For this reason, conduct that promises to advantage an individual on a given occasion may still be wrong if its performance conflicts with the standards essentially tied to the individual's goodness *qua* human being.

While neo-Aristotelian perfectionism may help to illuminate the non-optionality of moral requirements, it is less suited to explaining their inclusiveness. The approach generates standards of conduct for an individual on the basis of considerations about what it is for the individual to flourish or do well, as a member of its life-form. But standards of this kind do not necessarily have anything to do with the needs and interests of other persons. They define requirements of individual perfection rather than demands that flow from the problem of negotiating a social world of individuals whose interests count equally.

A reflection of this is the fact that some of the requirements of neo-Aristotelian ethics include demands that are entirely self-regarding and that have nothing to do with the effects of one's agency on the interests of other parties. One is under a requirement to adopt an optimistic attitude and to accept good things when they befall one[3] in just the sense in which one is under a requirement to be truthful to others or to uphold one's promissory commitments. Requirements of this kind, I would maintain, lack the essential connection to the interests of other parties that is characteristic of the modern conception of the moral right.

---

[2] See Philippa Foot, *Natural Goodness* (Oxford: Clarendon Press, 2001); also Michael Thompson, *Life and Action: Elementary Structures of Practice and Practical Thought* (Cambridge, MA: Harvard University Press, 2008).

[3] These examples are from Foot, *Natural Goodness*, p. 79.

4 Toward a relational account

In this section, I would like to sketch a different account of practical requirements, one that seems particularly illuminating in application to the moral case but that has wider implications as well. Let's return to the example of promissory commitments from Section 1. Promises clearly make a significant difference to the normative situation of promisors, and it is natural to ask how this difference comes about.

In the course of developing her natural goodness approach, Foot at one point takes up this question, observing that promises involve "a special kind of tool invented by humans for the better conduct of their lives, creating an *obligation*".[1] Promisors and promisees understand promises to bring into existence new and specially exigent normative facts, which we could refer to as obligations, and in virtue of their obtaining promisors bind themselves to fulfill a commitment to promisees.

Foot's basic proposal seems to invoke a new model of a practical requirement, distinct from the models considered previously, which we could call the relational paradigm. The promisor's thought, fully spelled out, might be that they owe it to the promisee to refrain from acting in ways at odds with the promissory commitment. Promising hence illustrates what we might call – to borrow the language of my recent book – a *normative nexus* between two different parties, the promisor and the promisee.[2] The promisor acquires, through the act of promising, a directed duty that is owed to the promisee, and the promisee acquires a corresponding entitlement or claim, against the promisor, to the fulfillment of this duty. This nexus explains the relations between the two from that point onward, so that if the promisor breaks their word, they do not merely act wrongly but wrong the promisee in particular.

A directed obligation of the kind at issue in this case intuitively has the kind of decisiveness that distinguishes deontic force from the merely recommending force of an ordinary enticing reason. This difference is reflected in the promisor's felt lack of deliberative discretion with respect to the reason the promise has created. The relational character of the duty acquired by the promisor illuminates this force: in any case exhibiting the sort of normative nexus in play with promising, the agent's reasons are essentially tied to someone's *claim* to the agent's compliance with the reason. The fact that the claim is held by someone other than the agent herself helps to explain why

[1] Foot, *Natural Goodness*, p. 51 (emphasis mine). Foot draws here on G. E. M. Anscombe, "Rules, Rights and Promises", as reprinted in Anscombe's, *Ethics, Religion, and Politics* (Minneapolis: University of Minnesota Press, 1981), pp. 92–103.

[2] See R. Jay Wallace, *The Moral Nexus* (Princeton: Princeton University Press, 2019), which was written at the same time as this contribution, and shares some root ideas (see especially chap. 2).

the agent cannot unilaterally decide to assign less than decisive weight to the reason it generates. The agent's reason in this case is part of a complex that also belongs to another party, and this helps us to understand why it should have a peremptory rather than a discretionary character.

As I noted earlier, it is also characteristic of practical requirements that they figure in deliberation as exclusionary reasons. They presumptively determine that some act is to be performed, taking the recommending features of incompatible actions out of consideration. Promissory commitments, however, are ordinarily understood to function in precisely this way. Having made a commitment to a promisee to do X, it should now seem to me irrelevant whether an incompatible action Y has much to recommend it. I owe it to my promisee to X, and have incurred a debt that must be repaid; against this background, the attractions of investing my efforts in something other than repayment of the debt become irrelevant (barring defeating circumstances). Directed obligations of this kind, which link two parties in a nexus of duties and corresponding claims, have for these reasons often been understood to represent the original notion of an obligation.[3]

But what about the inclusiveness of moral requirements – their connection with the equal standing of all moral persons, whose interests are taken to count equally in determining what each of us is obligated to do? The paradigmatic examples of relational duties considered so far do not seem well suited to accommodate this dimension of modern morality. Promissory commitments, after all, implicate only two parties in a nexus of directed duties and claims, the promisor and promisee. They rest on a causal transaction between the parties that are linked through them, in a way that does not seem to carry over to the broader category of moral requirements. We have duties to protect the interests of individuals who are liable to be affected adversely through our actions, regardless of whether we have previously undertaken a commitment to them to do so.

To extend the relational model to the entirety of the moral domain, we need to assume that a relational nexus can obtain between individuals who have not previously interacted with each other, in the ways that familiarly give rise to promissory or contractual commitments. We might hold, for instance, that there are things that we owe to other people just in virtue of the fact that they and we are equal members of the notional moral community, with interests that are susceptible to being affected by each other's agency. The result would be a universalistic conception of relational moral obligations as requirements that link us to other individuals in virtue of their equal standing as members of the moral community rather than on the basis of specific historical transactions we have entered into with them.

[3] See, for instance, Joel Feinberg, "The Nature and Value of Rights", *Journal of Value Inquiry* 4 (1970), pp. 243–257, at pp. 243–244; also Stephen Darwall, "Bipolar Obligation", as reprinted in his *Morality, Authority, and Law: Essays in Second-Personal Ethics*, vol. 1 (New York: Oxford University Press, 2013), pp. 20–39, at pp. 25–27.

It is not clear that this universalistic version of relational obligation can be worked out and defended, as a comprehensive account of "what we owe to each other". But it is a reason for taking this approach very seriously that it promises to illuminate the standing of moral rightness as a source of practical requirements, where this is something that is difficult to make sense of in terms of the other familiar models of a practical requirement.

5 Conclusion

I have canvassed several approaches to thinking about practical requirements and explored their implications for reflection about the nature of moral obligations. The relational approach has much to recommend it as a framework for understanding the special case of moral obligation, and one of my main aims has been to make the case for this conclusion. Stepping back from the vexed case of moral obligation, however, I believe that each of the models I have sketched has its place and that they characterize distinct potential bases for practical requirements that a given agent might be subject to.

Thus, it is a familiar point from discussions in moral theory that moral obligations sometimes seem to come into profound conflict with the personal projects and relationships that give our lives meaning and point. Consider Bernard Williams's re-imagining of the case of Gauguin, for instance, who faces the choice between pursuing his artistic calling in Tahiti or remaining in Paris to meet his obligations to his family.[1] It seems to me plausible to understand this as a case of conflict between practical requirements that have discrete sources: in what Gauguin owes to his family, on the one side, and in the demands of his identity as an artist on the other side.

Still further models might illuminate practical requirements of other kinds, including the coherence requirements that apparently constrain both theoretical and practical reasoning. A natural way of thinking about these requirements, to the extent they obtain, is to interpret them in voluntarist or identity-based terms. Thus, requirements to take the necessary means to our ends might be understood to involve commitments that we necessarily impose on ourselves when we undertake to pursue particular ends in the first place, and acknowledgement of requirements of coherence and consistency seems connected to our identity as rational believers.[2] But the relational model might have some application here, as well. Thus, the intrapersonal commitments through which coherence requirements are imposed might be understood to generate claims that we hold against ourselves, synchronically or diachronically, as agents and believers.[3]

In the end, I find the relational model especially promising when it comes to the obligations of morality. But I suspect we will need to draw on a plurality of models of a

[1] Bernard Williams, "Moral Luck", as reprinted in his *Moral Luck: Philosophical Essays, 1973–1980* (Cambridge: Cambridge University Press, 1981), pp. 20–39.

[2] See Christine M. Korsgaard, "The Normativity of Instrumental Reason", as reprinted in her *The Constitution of Agency*, pp. 26–68.

[3] Compare Sam Shpall, "Moral and Rational Commitment", *Philosophy and Phenomenological Research* 88(1) (2014), pp. 146–172.

practical requirement if we are to make sense of the full range and variety of practical requirements to which we are subject, and of the relations between them.[4]

[4] Many thanks to Kurt Sylvan for extremely valuable help in preparation of the final version of this chapter.

28. Normative pluralism and skepticism about 'ought *simpliciter*'

David Copp

There are different kinds of normative reason for action, including moral reasons and reasons of self-interest or prudence. There are also different kinds of normative requirement, including moral requirements and requirements of self-interest or prudence. Moreover, these can point us in different directions, or so I shall assume. Consider two examples:

> *Oxfam*: Suppose that Alice has some money that she could donate to Oxfam, but she could instead use it to upgrade her airline ticket to first-class. Oxfam would use the money to feed many hungry people. She has a morally decisive reason to donate to Oxfam, but she has a self-interested reason to upgrade her ticket, one that is decisive from the perspective of her self-interest. She asks herself, "What *ought* I to do? What ought I simply *to do*?"
>
> *Gyges*: Gyges considers whether to carry out his plot to assassinate the King, marry the Queen, and take over the kingdom. Carrying out the plot would help him achieve a life he views as wonderful, yet it would involve assassinating the King. Morally, he ought not to do this. There are morally decisive reasons not to do this, yet he has self-interested reasons to go ahead, reasons that are decisive from the perspective of his self-interest. He asks himself, "What *ought* I to do? What ought I simply *to do*?"

My goal is to explicate thoughts that Alice and Gyges might have in making their decisions, thoughts they might express by saying something like, "This is what I ought simply to do." What is the content of such thoughts? Are any such thoughts straightforwardly true?

One view, and perhaps the received view, is that, in saying, "This is what I ought simply to do," they might each be expressing their conclusion about what they ought to do "period," "finally," and "in the end," given the balance of reasons of all kinds that bear on the decision – and further, that if this conclusion were in fact true, this would genuinely *settle what they are to do period*, *simply*, and *in the end*. To facilitate my discussion, I will use the term "ought *simpliciter*" and stipulate that facts about what one ought *simpliciter* to do would genuinely *settle what one is to do period*, *simply*, and *in the end*. I will say that, according to the "received view," Alice and Gyges might each reach the conclusion, about one of the options they face, that this is what they ought *simpliciter* to do.

To evaluate the received view, it is necessary to consider the disagreement between two accounts of the nature of normative reasons and *oughts*, the "unified view" and the "pluralist view." The received view sits comfortably with the unified view but much less comfortably with the pluralist view. In the unified view, moral and self-interested *oughts* and reasons are based in different kinds of facts or considerations, but they are not fundamentally different in kind, and they are commensurable. Since

they are commensurable, the moral and self-interested *oughts* and reasons that bear on a person's decision can combine to yield a truth regarding what the person ought *simpliciter* to do. This might be so even in situations of the kind faced by Alice and Gyges, which I will call "conflict situations."

According to the pluralist view, however, normative reasons and *oughts* are all *relative to* some normative standpoint or domain, such as the standpoint of morality or that of self-interest.[5] As we will see, it is difficult in this view to avoid a "strong pluralism" according to which, in a conflict situation, there is no fact as to which of her eligible options a person ought *simpliciter* to choose.[6] For, on the pluralist view, oughts *simpliciter* would be relative to some normative standpoint, yet, by stipulation, they would genuinely *settle what to do period*, *simply*, and *in the end*. This would seem to mean that the normative standpoint that generates oughts *simpliciter* must have a kind of normative supremacy by comparison with the standpoints of morality and self-interest. As I will argue, however, it is doubtful that we can make sense of the required kind of supremacy. Indeed, it may seem, on the pluralist view, we lack a non-defective concept of an ought *simpliciter*. If so, it is doubtful that Alice and Gyges could even have *thoughts* as to what they ought *simpliciter* to do. Their thoughts must have some other content. Otherwise, their thoughts must be defective in some way – at best, they would fail to be true.

I begin, in section 1, by clarifying the distinction between the unified and pluralist views. In section 2, I defend my assumption that there can be conflict situations of the kind illustrated by my examples. In section 3, I argue for the pluralist view. In section 4, I respond to objections. In section 5, I present an argument from pluralism to the strong pluralist view that denies there is a supreme normative standpoint. In section 6, I draw a distinction that, I believe, defuses some worries about strong pluralism. In section 7, I rely on this distinction in responding to objections. There is also a brief conclusion.

[5] In place of the term, "self-interest," some would instead use "prudence." Here I discuss only *practical* reasons and *oughts*, which bear on actions, intentions, and the like. I set aside epistemic reasons. (For them, see Copp 2014.)

[6] Sagdahl (2013: 6) uses the term "pluralism" to refer to the position I am calling "strong pluralism." I distinguish the view that reasons and *oughts* are relative to standpoints, which I call "pluralism," from "strong pluralism," which adds the claim about oughts *simpliciter*.

1 Distinguishing the pluralist view from the unified view

In the pluralist view, normative reasons and *oughts* are in every case *relative to* some normative standpoint, such as morality or self-interest. Reasons are reasons-in-relation-to-a-standpoint and *oughts* are oughts-in-relation-to-a-standpoint. Normative reasons and *oughts* are like the weights of things in being essentially relational. In the unified view, however, there is one unified domain of reasons and *oughts* and, fundamentally, there is only one normative standpoint. In the unified view, moral reasons and reasons of self-interest are simply different kinds of reason-giving facts – morally relevant facts and facts about one's self-interest. Normative reasons and *oughts* are not relational to standpoints – as I shall say, they are "unqualified."[1] They are like the masses of things in not being (relevantly) relational. The unified view makes two main claims: There are *unqualified* reasons and *oughts*, which by stipulation are not relational to standpoints. And only *unqualified* reasons and *oughts* are normative. In the *pluralist* view, however, there are no unqualified *oughts* or reasons.

The unified view is widely shared by Kantians and neo-Kantians, as well as Humeans and neo-Humeans. Williams's view that there are only "internal" reasons is compatible with the unified view,[2] but so is the view that there are, in Williams's sense, "external" reasons (Williams 1981). The unified view is shared, as far as I can tell, by Broome (2013), Dancy (2006), Darwall (2006), Gauthier (1986), Korsgaard (1996), Parfit (2011), Scanlon (2014), Skorupski (2010), Smith (1994),[3] Wedgwood (2007), and Williams (1981), despite their differences on other issues. Sidgwick famously defended a pluralist view (1981 [1874]), and pluralist views have more recently been defended by Baker (2017), Chang (1997), Foot (1978), Finlay (2013), Killoren (2019), Sagdahl (2013), Tiffany (2007), and Wolf (1982).

The unified and pluralist views can agree that reasons are *facts*.[4] And they can agree that for a fact *to be* a reason is for it to stand in the "reason-relation" to appropriate relata. Scanlon describes the reason-relation as the "the relation of simply counting in favor of some action or attitude" (2014, 30). The reason-relation relates (1) a fact to

[1] This is a stipulation. One could use "unqualified reason" in other ways.

[2] A neo-Humean might say that all of a person's reasons are facts about how to further her ends. This is an example of a unified view. A pluralist can agree that *some* reasons are given by one's ends, that is, reasons-relative-to-one's-ends.

[3] Smith describes a pluralist view (1994: 95) but ultimately favors a unified view (ch. 6).

[4] See Dancy 2006: 137. Here I take a fact to be a true proposition.

(2) the person or persons for whom the fact is a reason and to (3) the action or action-type or other suitable relatum for which the fact is a reason.[5] The pluralist view holds, however, that one of the relata of the relation is, in each case, a normative standpoint. The unified view denies this. We can say the unified view takes the relation to be an "*n* place" relation whereas the pluralist view takes it to be an "*n+1* place" relation.[6]

In the unified view, various kinds of facts give us reasons, and all reasons are unqualified in the sense I explained. But this leaves room for a lot of disagreement among unifiers. They can disagree, for example, whether there are self-interested reasons, whether there are moral reasons, and which of these is normative.[7] I will largely ignore this kind of disagreement. In the Oxfam example, Alice plausibly has both a decisive self-interested reason to upgrade her ticket and a decisive moral reason to make the donation. Unified theorists can agree with this. I will address worries about conflict situations in the next section.

The pluralist view needs to explain what normative standpoints are. The basic idea is that a normative standpoint is a "source" of a kind of normative reason and requirement. Morality and self-interest are sources of moral and self-interested reasons and requirements, respectively, so they count as normative standpoints. There may be other such standpoints.

One might question whether self-interest and morality are well enough delineated to serve as relata of the reason-relation. There are, however, many examples of relations that take vague or contested entities as relata. Biological species are not precisely delineated from one another, yet biologists treat gorillas and orangutans as different species. The idea of a reason-relation that takes normative standpoints as relata seems to be no more problematic than the idea of species membership. In any case, moreover, many philosophers have discussed the moral "point of view" (e.g. Baier 1958), and many philosophers refer to self-interest or prudence (see, e.g., Gauthier 1986). Perhaps it is a mistake to think that morality and self-interest can be referred to in this way, but we would need an argument to show this. Furthermore, it is open to unifiers to agree that morality and self-interest are sources of reasons and requirements. The unified view can agree that morality and self-interest are in *this* sense different normative standpoints.

Unfortunately, the term "normative" is a piece of philosopher's jargon, and different philosophers use it differently. A currently widely shared approach is to explain normativity in terms of reasons.[8] The problem with this idea is that there are reasons and *oughts* that are not normative in the philosophically important sense, such as reasons in games and, perhaps, reasons of etiquette. There is a reason in chess to protect one's

[5] There may be additional relata, such as situation-types. See Dancy 2006: 137; Skorupski 2010: 35–37; Scanlon 2014: 30.

[6] I assume that the unified view denies that its single fundamental normative standpoint is a relatum of the reason-relation.

[7] Tristram McPherson reminded me of this.

[8] Parfit distinguishes four conceptions of normativity and contends that one of the four, normativity in the "reason-implying sense," is the philosophically most important (2011, II: 267–269).

queen, and one ought (chess-wise) to do so. But, *in themselves*, this reason and this *ought* have no authority over anyone. They are not in themselves normative.

We can gesture at the idea of normativity by relating it to the idea of a mistake. Suppose that a normative reason or *ought* of some kind K favors a person's doing A. If she knows this, but decides what to do without taking it into account, and treating it as counting in favor of doing A, she is thereby making a normatively significant mistake.[9] Now, intuitively, a person who ignores reasons and *oughts* of morality and self-interest does thereby make a significant mistake, other things being equal. But a person who ignores a game-reason, such as "chess-reason," is not thereby making a significant mistake, not unless there is a moral or self-interested reason or a normative reason of some other kind to take her chess-reason into account. Intuitively, morality and self-interest are not games.[10] Unlike games, they are normative and authoritative. They address significant problems of normative governance that would be faced by people in all realistic situations (see Copp 2009).

Intuitively, then, there is a significant difference between the reasons and *oughts* provided by games, which are not in themselves normative, and the reasons and *oughts* of morality and self-interest, which are normative. Unifiers and pluralists agree, I assume, that there is this difference, but they provide different accounts of what the difference comes to. In the unified view as I understand it, the reasons and *oughts* of morality and self-interest are fundamentally unqualified, and only unqualified reasons and *oughts* are normative. The reasons of games are not unqualified. In the pluralist view, normative reasons and *oughts* are all relativized to some normative standpoint or other, but the "standpoint" of a game is not a normative standpoint. Hence game-reasons are not normative.

Accordingly, both views face a challenge. Pluralists need to explain what distinguishes normative standpoints, such as the standpoints of morality and self-interest, from the standpoints of games (Tiffany 2007; Dorsey 2013). Normative standpoints have a property we might call the N property, and the standpoint of a game lacks this property. I have provided an account of this property in other work (esp. Copp 2009; also Copp 1995; Copp 2007a). It is beyond the scope of this chapter to explain my account here. For our purposes, it does not matter what the N property is. The important claim is that a plurality of standpoints have the N property, but games do not. Unifiers face a corresponding challenge. They need to explain what distinguishes normative reasons and *oughts* from the reasons and *oughts* of games. Normative reasons and *oughts* have a property that reasons and *oughts* of games do not have, a property we might call the N* property. Perhaps unifiers will say this is the property

[9] From the pluralist perspective, this would be a mistake of kind K, just as there are moral mistakes and prudential mistakes. There are also "chess-mistakes," but, intuitively, chess-mistakes are not normatively significant in the way that moral and prudential mistakes are.

[10] Tiffany proposes "deflationary pluralism," according to which moral reasons are on a par with reasons in games (2007). This position raises the challenge I go on to address.

of being unqualified. If so, they need to explain why the fact that a reason or *ought* is unqualified ensures that it is normative.

Pluralists and unifiers give different answers to the question whether there are oughts *simpliciter*. Unifiers say that what one ought *simpliciter* to do is the action that is best supported by the relevant unqualified reasons. Pluralists object that there are no unqualified reasons. Despite this, however, pluralism is compatible with the existence of oughts *simpliciter* in cases in which all relevant normative standpoints agree – cases in which an action A is such that, for every relevant normative standpoint K, there is most reason of kind K to A.[11] Recall that, as I stipulated, the idea that there is something that someone ought *simpliciter* to do is the idea that there is something she ought to do where this fact genuinely *settles what she is to do period*, *simply*, and *in the end*. (It is not the idea of an *ought* that is not relational to a standpoint.) If some action A is such that, for every relevant normative standpoint K, there is most reason of kind K to A, then, pluralists should say, this settles what to do, period, simply, and in the end. In such cases, pluralists should say, a person ought *simpliciter* to do what is most supported by reasons of all relevant kinds.

Pluralism would be compatible with the existence of *oughts simpliciter* in a wider range of cases, including conflict situations, if it could make sense of the idea that there is a standpoint that is special – a standpoint that is normatively "supreme" – one that determines what a person simply ought to do, period, and that settles what to do. Call this hypothetical standpoint the Supreme standpoint. This standpoint would determine definitively the relative degrees of importance of reasons and *oughts* of all kinds. It would determine, for example, whether morality is normatively "overriding." For all we have seen, the existence of such a standpoint might be compatible with the pluralist view (see Copp 2009). If there is *no* such standpoint, however, then, in the pluralist view, an agent in a conflict situation simply faces an array of standpoint-relative normative verdicts, none of which determines what she ought *simpliciter* to do.

[11] Or better: for some standpoint K, there is most reason of kind K to A, and for no standpoint K* is there more reason of kind K* not to do A than to do A. No standpoint speaks all-in against A. I am here ignoring some complications. See Sagdahl 2013: 37.

2 Can there be conflict situations?

I assume there can be conflict situations of the kinds illustrated by the examples of Gyges and Alice. These are situations in which certain facts give an agent decisive self-interested reason to do (or not to do) something and certain facts give the agent morally decisive reason not to do (or to do) that thing. Some versions of the unified view would agree with this assumption, but other versions would reject it. Some would claim that morally decisive reasons, such as Gyges's reason not to murder the king, extinguish or perhaps subsume any putatively conflicting reasons of self-interest. Others would claim that there are only reasons of self-interest, that there are no moral reasons. I cannot adequately address these views here, but I find them completely implausible, for reasons I will briefly explain.

Consider first the idea that decisive moral reasons *extinguish* or *eliminate* any putatively conflicting self-interested reasons (see McDowell 1978: 26, 29). In this view, Gyges's moral reason not to murder the king eliminates any self-interested reason he might have had to carry out his plot. The power and love he would gain by carrying out the plot give him no reason at all to go ahead. I find this line of thought completely implausible. Similarly, in the Oxfam example, I find it completely implausible that the comfort Alice would enjoy if she were to upgrade her ticket is no reason at all for her to upgrade her ticket. She is, after all, under no obligation to Oxfam, and she is entitled to use the money to upgrade the ticket. These examples illustrate, then, I think, the implausibility of the idea that decisive moral reasons eliminate putatively conflicting self-interested reasons. This leaves it open, of course, that moral reasons always "outweigh" any conflicting self-interested reasons.

One might suppose instead, as welfarist consequentialism suggests, that moral reasons *subsume* self-interested reasons (see Parfit 1984). In this view, everyone's welfare is taken into account by morality, which is concerned with enhancing the *general* welfare. Alice does have reason to enhance her welfare but only insofar as, and in virtue of the fact that, doing so enhances the general welfare. This line of thought also seems implausible, even if we do not question consequentialism. Suppose that Alice and Bill have equally painful headaches, and suppose Alice has the only available dose of analgesic. She could use it relieve *her* headache or she could instead give it to Bill. Suppose further that, if Alice uses the pill to relieve her headache, two friends of Bill will develop intense headaches of their own, in sympathy for Bill. Intuitively, Alice still has

the self-interested reason to help herself even though helping herself would not enhance the general welfare. The subsumption view implausibly denies this.[1]

Consider now the different idea that all reasons are grounded in self-interest, or in the agent's advantage or welfare (see Gauthier 1986). In this view, Gyges has no reason not to murder the king, and Alice has no reason to make the donation, unless these courses of action somehow redound, respectively, to Gyges' or Alice's own benefit. This view, I submit, is completely implausible. One might think it is supported by Williams's argument (1981) that there are only "internal reasons" – reasons grounded in the agent's "subjective motivational set." But his arguments give us no reason to deny that there are moral reasons. Williams claims that if a person has a reason to do something, she must be capable of doing the thing *for* that reason (1981: 106–107). If she were to do it for that reason, says Williams, she would have to be motivated to do so, so she would have an internal reason to do it (107). As Williams admits, however, it does not follow that there are no external reasons (108), nor does it follow that there are no moral reasons (see Korsgaard 1986). It is obvious that if Gyges were to refrain from murdering the king for a moral reason, he would have to be appropriately motivated. But it does not follow that the reason is *grounded* in his motivations, nor that, in order for the reason to exist, Gyges must be motivated by it. All that follows is that, in order for the reason to exist, Gyges must be *capable* of being motivated by it, and that for Gyges to act on the reason, he must be motivated appropriately when he acts.

Some versions of the unified view would agree that there are decisive moral reasons for Gyges not to carry out his plot and decisive self-interested reasons for him to go ahead with it.[2] According to these unified views, either the moral reasons that support Alice's making the donation are equally balanced or on a par with the self-interested reasons that support upgrading her ticket, or one of these options is such that she ought *simpliciter* to do it (see Chang 2002). Both of these alternatives depend on there being a definitive ranking of the relative importance of moral and self-interested reasons that settles the relative weight of the different reasons bearing on Alice's decision, and that settles this *period*, and not relative to some particular standpoint. We might not know what the definitive ranking is, but the unified view seems committed to the view that there is such a ranking. This seems to me to be wishful thinking.

Moreover, the unified view's account of reasons strikes me as puzzling. Understood in one way, it holds that there is only one normative standpoint – call it the "vanilla" standpoint – and that it determines what reasons there are.[3] So understood, unifiers

[1] I am grateful to Andrew Forcehimes for helpful discussion of the subsumption view.

[2] My assumption that there are conflict situations does not beg the question.

[3] On the unified view, the vanilla standpoint determines the extension of the reason-relation. Accordingly, we can jerry-rig an *n+1* place reason*-relation by combining this determination relation with the *n* place reason-relation. The vanilla standpoint would be a relatum of this relation. There is still a difference between the unified and pluralist views because the unified view would say that the vanilla standpoint is the *only* standpoint that is a relatum of the reason*-relation.

and pluralists share the idea that reasons are grounded in normative standpoints but unifiers hold that there is only the one such standpoint. In this understanding, the role of the vanilla standpoint is analogous to, but more critical than, the role that would be played in the pluralist view by the Supreme standpoint. For in the unified view as we are understanding it, if there is no vanilla standpoint, there are no unqualified reasons, and this means there are no normative reasons at all. In a different and perhaps better understanding, the unified view says that unqualified reasons are grounded independently of any standpoint at all. But it is not clear how to explain this.

To my mind, pluralism provides intuitively plausible accounts of Gyges's and Alice's cases. It says that there are decisive reasons for Alice to make the donation to Oxfam, and there are decisive reasons for her to upgrade her airline ticket, but these are reasons of different kinds, relating to different normative standpoints. Assume that the moral and the self-interested standpoints agree in ruling out all but these two options – her "eligible options." In this case, the pluralist can say that Alice ought *simpliciter not* do anything other than either make the donation or upgrade her ticket, but, assuming there is no supreme standpoint, it says that whether either of these options is such that she ought *simpliciter* to do it depends on whether, in the end, morality and self-interest agree. This strikes me as an intuitive picture of Alice's situation. Intuitively, I think, morality and self-interest are distinctive normative standpoints and, although each is a source of reasons, these reasons do not all count in the same way, and they might not be commensurable.

It is intuitive, I think, that there are different kinds of normative reasons and *oughts* including self-interested ones and moral ones. Intuitively, a person makes a significant mistake if she is aware of considerations of these kinds yet ignores them in deciding what to do. Moreover, intuitively, there is no way to gauge the relative weight or importance of different kinds of consideration except by evaluating the matter from some normative standpoint. So, I think, in many ways, the pluralist view is intuitively very attractive. In other ways, as we will see, the view seems to run up against our intuitions.

3 An argument for the pluralist view

In the pluralist view, reasons are in every case *relative to* a normative standpoint. To see how one might come to believe this, it will help to consider prosaic cases where we think that a property is relational even though the fact that it is relational is not immediately apparent. Weight is an example. We might simply say that a rock weighs ten pounds, yet weight is a relation between an object and a gravitational field. Without any change in its intrinsic properties, an object can have different weights in different gravitational fields. We have a relational view of a "feature" of things if we realize there is a relevant "parameter" such that, even if there is no change in a thing's intrinsic properties, the thing can have or not have that feature depending on the value of the parameter. A rock can have a weight in one gravitational field that it does not have in a different one.

The example suggests a way to decide between the pluralist and unified views. We can ask whether a fact can be a reason for an agent *to do* something and also *not* be a reason for that agent to do that thing or be a reason for that agent *not to do* that thing. If this is possible, it suggests there is a missing parameter such that being-a-reason is relational with respect to that parameter.

Gyges plots to assassinate the King, to marry the Queen, and to take over the kingdom; he values power and love, and carrying out his plot would help him to achieve a life of power and love at small risk to himself. Call this the "Gyges fact." It is a decisive self-interested reason for him to carry out his plot, yet of course it is *not a moral reason* for him to carry out the plot. The unified view can agree with this, but on the unified view, all reasons are fundamentally of one kind. "Moral reasons" and "self-interested reasons" are different kinds of facts, but fundamentally, they are simply reasons, for, in the unified view, morality and self-interest are not parameters of the reason relation. This means that we can drop the modifiers and restate our conclusion.[1] We can say that the Gyges fact is both a decisive reason for Gyges to carry out his plot and not a reason for him to do this. In the unified view, the Gyges fact both stands in the reason-relation to Gyges's carrying out his plot and it does not stand in this relation to his carrying out his plot. This is the first step in the argument.

Before going any further, it will be useful to consider two objections. First, the unified view could deny that Gyges fact is a self-interested reason for Gyges to carry

[1] Ruth Chang helped me here.

out the plot. The problem, however, is that the Gyges fact shows that Gyges's carrying out of his plot will most likely contribute to his achieving the kind of life he values and that he could not achieve in any other way. This seems to be a paradigm example of a self-interested reason.

A second and better objection would begin by reminding us that, in the unified view, a moral reason is a fact that is both morally relevant and a reason. So even though the Gyges fact is a self-interested reason, but not a moral reason, for Gyges to carry out the plot, it does not follow, in the unified view, that the Gyges fact is both a reason and not a reason. Assuming it is not a moral reason, all that follows is that either it is not a reason or it is not a fact of a morally relevant kind. So, assuming the Gyges fact is a self-interested reason to carry out the plot, the unified view should say that it is not morally relevant. The problem, however, is that the Gyges fact is plainly morally relevant because it shows that Gyges's plan is to assassinate someone simply for personal gain, and this is not morally acceptable. Indeed, the Gyges fact seems to be a moral reason for Gyges *not* to carry out his plot, so it seems obviously to be morally relevant.

The upshot is that the unified view cannot avoid the first step of my argument in either of these ways unless it gives a highly counter-intuitive account of the Gyges example. It could implausibly deny that the Gyges fact is a self-interested reason for Gyges to carry out the plot, or it could implausibly deny that the Gyges fact is morally relevant. Otherwise, it is committed to accepting that the Gyges fact is both a reason for Gyges to carry out the plot and not a reason for him to do this. This seems incoherent.

The next step in the argument invokes the following key premise. A fact <F> cannot be a decisive normative reason for A and also *not* be a normative reason for A (or be a decisive normative reason for *not-A*) – not unless there is a relevant parameter such that, given one "value" for the parameter, <F> is a normative reason for A, and given a different value for the parameter, <F> is not a normative reason for A (or is a normative reason for not-A). Why? It is platitudinous that a fact is a normative reason for some action when it counts in favor of that action. A given fact <F> cannot count in favor of doing A and also *not* count in its favor (or count in favor of *not* doing A) – not on the same occasion, for the same agent, and where the circumstances are the same. It cannot do so unless there is some "respect" in which <F> counts in favor and a different respect in which <F> does not count in favor (or counts against) doing A. Hence, normative reasons are relational to "respects." Or, we could say, the reason-relation holds between facts, persons, act-types, and "respects" – and there may be other relata as well.

The key premise rests in part on the idea is that a fact is a reason to do something if and only if (and because) it stands in the *reason-relation* to a set of relata. The reason-relation relates (1) a fact, (2) the person or persons for whom the fact is a reason, (3) the action (or other suitable relatum) for which the fact is a reason, (4) the circumstances that are relevant, and perhaps some additional relata. Given this,

it follows that if a fact is a reason for an agent to do A, and is also not a reason for the agent to do A (or is a reason for the agent not to do A), there has to be some difference in the relata of the reason-relation that explains this. In the Gyges example, the Gyges fact is both a reason for Gyges to carry out his plot and not a reason for Gyges to carry out his plot, and this is so despite the fact that the agent, action, and circumstances are held constant. Some other relatum must be involved to explain this.

But if the reason-relation relates facts, persons, act-types, circumstances, and *respects*, what are these respects? The examples suggest something quite obvious. In relation to Gyges's self-interest, the Gyges fact is a reason for him to carry out his plot, but in relation to morality, it is no reason for Gyges to carry out his plot. We have been led to the idea that these respects are "normative standpoints" such as self-interest or morality. It is hard to resist this idea. Self-interest and morality are the most plausible candidates for the relevant respects in the Gyges example. A similar argument could be given for the Oxfam case.

The normative situation in the Gyges example is completely coherent in the pluralist view. The reason-relation relates the Gyges fact to Gyges' carrying out the plot relative to the standpoint of self-interest yet it does not relate the Gyges fact to Gyges's carrying out the plot relative to the moral standpoint.

We have, then, an argument for pluralism. It starts from the observation that, on the same occasion, for the same agent, and where the circumstances are the same, a given fact <F> can be a decisive self-interested reason for the agent to do A but not be a moral reason to do A, or it can be a decisive moral reason to do A and a decisive self-interested reason not to do A. It then contends that this is coherent only on the assumption that there is some relatum of the reason-relation such that <F> counts in favor of doing A relative to one "value" for this relatum but does not count in favor of doing A, or counts against doing A relative to a different value for this relatum. In the Gyges example, these values were the standpoints of morality and self-interest. To generalize, reasons are relative to normative standpoints, such as morality and self-interest.

Turn now to *oughts*. A similar argument shows that *oughts* must also be relativized to standpoints. In virtue of the Gyges fact, Gyges ought in his self-interest, all-in, to carry out his plot, yet of course it is not the case that, in virtue of the Gyges fact, Gyges ought *morally*, all-in, to carry out his plot. In the unified view, since "moral oughts" and "self-interested oughts" are simply *oughts*, we can drop the modifiers and restate our conclusion. We can say that, in virtue of the Gyges fact, Gyges ought all-in to carry out his plot, and it is not the case that, in virtue of the Gyges fact, Gyges ought all-in to carry out his plot. This seems incoherent. A given fact <F> cannot make it the case that a person ought all-in to do A and also *not* make it the case that the person ought all-in to do A – not on the same occasion, for the same agent, and where the circumstances are the same. It cannot do so unless there is some "respect" in which <F> makes it be the case that the person ought to do A and a different respect in which <F> does not make this be the case. Hence, *oughts* are relational to

"respects." And as before, morality and self-interest are the most plausible candidates for the relevant respects

4 Objections

The crucial premise in the argument is the claim that a fact <F> cannot be a decisive normative reason for A and also not be a normative reason for A (or be a decisive normative reason for not-A) – not unless there is a relevant parameter such that, given one value for the parameter, <F> is a decisive reason for A, and given a different value for the parameter, <F> is not a reason for A (or is a decisive reason for not-A). One might think there are counterexamples.[1]

First, suppose Aurelia's doing A will save many people but harm a few. Call this the "A-fact." It is a strong moral reason for Aurelia to do A but also a weak moral reason for her not to do A. Yet there are not different moral standpoints such that the A-fact counts in favor of Aurelia's doing A relative to one but counts against relative to another. Second, suppose that Barbara's doing B would satisfy an important desire of hers but frustrate another important desire. Call this the "B-fact." In a familiar view, this fact would be a self-interested reason for Barbara to do B but also a self-interested reason for her not to do B. To explain this, we do not need to distinguish different values of some parameter of the reason-relation.

The simplest response to both examples is to argue that it is different "components" of the A-fact and the B-fact, not these facts themselves, that point in the different directions. It is important for our purposes to be strict as to which fact stands in the reason-relation to a given action. When a fact stands in the reason-relation to an action, we can sometimes augment the fact with additional detail such that the resulting more complex fact is *not* a reason for the action. The fact that I want to eat this sandwich might be a reason to eat it, but the complex fact that [I want to eat this sandwich and it has been poisoned] is not a reason to eat it. It would be a mistake to suppose that the complex fact that [I want to eat this sandwich and it has been poisoned] is *both* a reason to eat it *and* not a reason to eat it. Borrowing a term from logic, we can say that the reason-relation is "non-monotonic." This point helps us to understand the putative counter-examples.

In Aurelia's case, we should say that the fact that stands in the reason-relation to Aurelia's doing A (in relation to morality) is the fact that by doing A she will save many people. The fact that stands in the reason-relation to her not doing A is the fact that by doing A she would harm some people. The conjunction of these facts is the A-fact as a whole, but it does not follow that the A-fact as a whole is both a moral reason for Aurelia to do A and a moral reason for her not to do A. For the

[1] Tristram McPherson pressed this point and proposed the following examples.

reason-relation is non-monotonic. Of course, the A-fact as a whole entails the fact that is a reason for Aurelia to do A, and it also entails the different fact that is a reason for her not to do A. It entails but is distinct from these facts.

A similar response is available in the desire example. The fact that stands in the reason-relation to Barbara's doing B (relative to self-interest) is the fact that doing B would satisfy an important desire. The fact that stands in the reason-relation to Barbara's not doing B is the fact that doing B would frustrate another important desire. It does not follow that the B-fact as a whole is both a self-interested reason for Barbara to do B and a self-interested reason for her not to do B. For the reason-relation is non-monotonic.

One might now contend that a similar move can be used to undermine my claim that the Gyges fact is both a self-interested reason for Gyges to go ahead with his plot and no moral reason for Gyges to do that thing. One might contend that it is different components of the Gyges fact, not the Gyges fact itself, that point in different directions. This, however, is a mistake. The Gyges fact is that carrying out his plot to assassinate the king and so on would help Gyges achieve the life he values at low risk to himself. This fact is not a moral reason for Gyges to carry out the plot, partly because the plot is to assassinate someone and partly because Gyges's motive is merely personal gain. The Gyges fact is, however, a self-interested reason for Gyges to carry out the plot, partly because it includes the fact that he stands to gain the life he values, but partly as well, in the case as I am imagining it, because the risk to him of carrying out the plot is small. The Gyges fact as a whole speaks in favor of the plot from the standpoint of Gyges's self-interest but speaks against the plot from the standpoint of morality.

It seems to me, then, that both of the key premises of the argument are secure. A fact can be a self-interested reason to do A but not be moral reason to do A. If so, then there is a relevant parameter of the reason-relation such that, given one "value" for the parameter, the fact is a reason to do A, but given another value, the fact is not a reason to do A.

5 Oughts *simpliciter* and the strong pluralist view

In this section, I provide an argument for the proposition that, assuming pluralism, there are no oughts *simpliciter* in conflict situations.[1] As I explained, pluralism holds that there are oughts *simpliciter* in conflict situations only if there is a supreme normative standpoint – a standpoint the verdicts of which determine what a person simply ought to do, where this settles what she is to do period, and in the end. Pluralism is perhaps compatible with the existence of such a standpoint, and in Copp 2009, I suggested a strategy a pluralist could use to ground the idea that there is a supreme normative standpoint. But the argument I am about to present undermines the thesis that there is such a thing. The pluralist view is a premise of the argument.[2]

Consider a normative standpoint S and assume it is the Supreme one. For now, set aside worries about its content. Like the moral and self-interested standpoints, S presumably would take all relevant facts and normative considerations and all normative reasons and *oughts* into account before rendering a verdict as to what a person ought-S to do "all-in," or all things considered (see Baker 2017, sec 2.2). (In what follows, I will often omit indicating that we are discussing all-in *oughts*.) But S is distinguished from other normative standpoints on the basis that, by stipulation, its verdicts as to what a person ought all-in to do determine what this person ought *simpliciter* to do. And since *oughts* imply reasons, if a person all-in ought *simpliciter* to do something, she has an all-in reason so to act. Let me stipulate that such reasons count as *conclusive*, since one ought *simpliciter*, all-in, to act in accord with them. Since S is supreme, if a person all-in ought-S to do A, she ought *simpliciter* to do A, and she has a conclusive reason to do it.

We are assuming that S is a normative standpoint, so we are, I think, committed to there being some fact about S that grounds its normativity. Similarly, I think, there is some fact that grounds the normativity of morality and some fact that grounds the normativity of self-interest. I have proposed a unified theory of such grounding

[1] As I explained, pluralism allows that there can be oughts *simpliciter* in situations in which all relevant normative standpoints recommend the same action.

[2] In an earlier paper (Copp 2007c), I proposed a different argument, which has been criticized by McCleod (2001), Tiffany (2007), Dorsey (2013), Sagdahl (2013), and Baker (2017). One important difference between that earlier argument and my current argument is that, here, I make it explicit that normative pluralism is a premise in the argument, and, of course, I have provided an argument for pluralism.

facts in other places, but the details are not relevant here (see Copp 2009, 2015). The important point is that, in addition to the fact about S that grounds its normativity, there must be some fact about it that grounds its supremacy. For if S is the Supreme standpoint – if it determines what one ought *simpliciter* to do – it follows trivially that it has some property that is not possessed by other normative standpoints. Call this the property of supremacy. Supremacy is a normative property, and I assume that the normative properties of a thing must be grounded by or explained by other facts about the thing that distinguish it from things that lack these properties. So the supremacy of S must be grounded in some fact about S. Call this fact the "supremacy grounding fact" or "SG fact."

Notice that the argument rests on the assumption that the normative properties of a thing must be grounded by, or explained by, other facts about the thing. So if some standpoint S is supreme, this fact must be grounded in some other fact about S. If morality is supreme – if it is "overriding" – this must be grounded in some other fact about morality. This assumption could be challenged, and I will return to it after presenting the argument.

To make my argument concrete, it will help to assume that the supreme standpoint S is identical with the moral standpoint. Nothing in the argument depends on this assumption. We could instead assume that S is identical with the self-interested standpoint or with some third standpoint. When nothing will be lost by abbreviation, I will revert to the label "S." But when it will help with explaining the issues, I will use the assumption that S is the moral standpoint.

Morality and self-interest are both normative standpoints, so, I assume there is some fact about each of them that grounds its normativity. In virtue of these grounding facts, there are moral reasons and *oughts* as well as self-interested reasons and *oughts*. But on our assumption that S is the moral standpoint, morality has the supremacy property. It is overriding. And this fact must be grounded in some other fact about morality, the SG fact, that distinguishes it from self-interest. It is in virtue of the SG fact that there is an all-in *conclusive* reason for Alice to do what she morally ought all-in to do rather than what she all-in ought in her self-interest to do.

To see what would be involved in morality's supremacy, consider the claim that Alice ought morally all-in to donate. It follows that Alice ought morally to do donate in preference to anything she ought-X to do, where X is any standpoint that disagrees with morality about what to do. But this does not distinguish morality from self-interest. For it is also the case that Alice ought in her self-interest, all-in, to upgrade her ticket. And it follows that she ought in her self-interest to upgrade her ticket in preference to anything she ought-Y to do, where Y is any standpoint that disagrees with self-interest about what she is to do. Moral *oughts* and self-interested *oughts* are, in this sense, *exclusionary*. But *oughts-S* are not merely exclusionary. For, by stipulation, S is the supreme standpoint, which means that S determines what one ought *simpliciter* to do, and, by stipulation, this means that what one all-in ought-S to do *settles* what one is to do. So if morality is supreme or overriding, as I am assuming, then Alice

ought *simpliciter* to donate rather than anything she ought-X to do. If Alice ought *simpliciter* all-in to donate, it follows by stipulation that she has a conclusive reason to donate – since *oughts* imply reasons. Let us think about this *ought* and the nature of the implied all-in conclusive reason.

Alice can ask: "I see that I ought morally to make the donation rather than to upgrade my airline ticket. And I see that I ought in my self-interest to upgrade my ticket rather than make the donation. But why does the fact that I ought morally to donate *settle* what I ought to do, *finally and in the end*? There must be a special, *conclusive* reason to do what I ought morally to do rather than what I ought in my self-interest to do. What is this reason?" If morality is supreme, there would have be such a reason, for, otherwise, any moral reason for her to do what she ought morally to do rather than what she ought in her self-interest to do would have the same status as the self-interested reason she has to do what she ought in her self-interest to do rather than what she ought morally to do. Our answer to Alice's question cannot be that she ought morally to donate since she similarly ought in her self-interest not to donate. And our answer cannot be that she has a decisive moral reason to donate since she similarly has a decisive self-interested reason not to donate. What she is looking for is an all-in *conclusive* reason to do what she ought morally to do, rather than what she ought-X to do, for *any* standpoint X that disagrees with morality.

So, on my assumption that morality is supreme, and that its supremacy is grounded in the SG fact, Alice has a conclusive all-in reason to do what she ought morally to do rather than what she ought-X to do. This reason could be the bare fact that morality is supreme, or it could be the SG fact, which grounds the supremacy of morality. For now, let us assume that the SG fact would be the reason. That is, assume that the SG fact is the all-in conclusive reason for Alice to do what she all-in ought morally to do rather than what she all-in ought-X to do.

The argument is heading toward the conclusion that, on the assumption that morality is supreme, Alice has both a conclusive "first-order" or "*de re*" moral reason [to make the donation rather than not], and the further conclusive "second-order" or "*de dicto*" reason [to do what she ought morally to do rather than what she ought-X to do, for any standpoint X that disagrees with morality].[3] To challenge my argument, one might object that there is no such second-order reason.[4] One might contend that the property of normative supremacy is primitive and unanalyzable.[5] And so, one might suggest, to explain the supremacy of morality, there is no need to postulate a second-order or *de dicto* reason [to do whatever one ought morally to do rather than (e.g.) whatever

[3] Peter Shiu-Hwa Tsu helped me here.

[4] This objection was suggested by Tristram McPherson. McPherson worried that the argument might depend on what Schroeder calls "the Standard Model" (2007: 43), but it does not. I leave it to the reader to see why. I also leave it to the reader to see how the argument could be run without assuming there is this second-order reason. I think there is such a reason, and assuming there is helps to simplify the argument.

[5] In discussing my argument in Copp 2007c, McCleod suggests a move of this kind (2001: 274, 286).

is in one's self-interest]. It is enough that Alice have the conclusive first-order or *de re* moral reason [to make the donation]. My response is that, even if the property of normative superiority is primitive and unanalyzable, it does not follow that morality might be supreme even if there is no such second-order reason.

For a deeper response, I want to argue that, assuming morality is supreme, there are both the first-order and second-order reasons, and both are conclusive. First, if morality is supreme, it follows by stipulation that Alice's first-order all-in moral reasons are conclusive. For example, if morality is supreme, then Alice's all-in first-order moral reason to donate the money is conclusive, since Alice ought *simpliciter* to act in accord with it. I am suggesting that its being conclusive must be explained by some further fact, such as the SG fact. Second, if morality is supreme, Alice also has an all-in conclusive reason [to do whatever she has all-in moral reason to do rather than what she has all-in X-reason to do, for any standpoint X that disagrees with morality]. For she ought *simpliciter* to do [what she has all-in moral reason to do rather than what she has all-in X-reason to do]. The existence of the all-in conclusive reason follows. I call it "second-order" because it concerns the relative weight of moral reasons and X-reasons. On our assumptions, then, the SG fact is an all-in conclusive second-order reason for anyone [to do whatever she has most all-in moral reason to do, rather than what she has most all-in X-reason to do], and [to do whatever she all-in ought morally to do, rather than what she all-in ought-X to do, for any standpoint X that disagrees with morality as to what the person ought to do].

Next, consider this second-order reason. Let me revert to using "S" to refer to the supreme standpoint. According to pluralism, the reason to do what one ought-S to do rather than what one ought-X to do must be relative to some standpoint or other. Call this standpoint R. The SG fact gives Alice a normatively conclusive all-in R-reason to do what she ought-S to do. Further R must be either identical to S or distinct from S.[6]

We need to deny that R is distinct from S. There are reasons of theoretical simplicity to deny this, but also, I think, we need to deny it in order to avoid regress or circularity. To see this, assume that Alice's (relevant) reason to do what she ought-S to do is an R-reason, where R is distinct from S. This reason must be normatively conclusive. It must take normative priority over reasons of any other kind X that Alice might have to do something else. But now Alice can ask, "What is the conclusive reason for me to do what I have R-reason to do rather than what I have X-reason to do?" The answer will refer to some fact, and *this* fact must be a conclusive reason for Alice to do what she has R-reason to do rather than what she has X-reason to do. This reason must, in turn, be relative to some standpoint Q. If Q is distinct both from R and from S, we

[6] This argument is different from the argument in Copp 2007c. There, my key move was to claim that the Supreme standpoint must be supreme as assessed from the perspective of some normative standpoint, which would itself have to be supreme. McCleod criticized this move (2001: 286–287). Here, the key move is different. It is to claim that the reason to do what one ought-S to do must be a reason in relation to some normative standpoint. This move relies on pluralism, which I have defended here.

are embarked on a vicious regress.[7] If Q is identical to S, we have a vicious circularity. We plainly have entered a morass it is better to avoid and the way to avoid it is to deny that R is distinct from S.

Suppose, then, that R is identical to S. We reached the conclusion that the SG fact is the all-in conclusive reason for Alice to do what she ought-S to do rather than what she ought-X to do. We are now supposing that this reason is an S-reason. And since it is normatively conclusive, it takes normative priority over reasons of other kinds that Alice might have to do something else. It settles what Alice is to do. Alice ought *simpliciter* to do what she has all-in conclusive S-reason to do – or to do what she all-in ought-S to do. The trouble is that our account now runs in a tight circle. Let me spell it out. It says that the SG fact grounds the fact that the SG fact is the conclusive S-reason to do what one all-in ought-S to do. The SG fact certifies itself.

Assume again that morality is the supreme standpoint. Alice says: "What is the all-in conclusive reason for me to do what I ought morally to do rather than what I ought-X to do, for any standpoint X that differs with morality?" Our response is that the SG fact is this reason. She can now ask, "Why is this fact an all-in *conclusive* reason?" Our answer is that the SG fact is a conclusive reason because of the SG fact. We can say this using more words. We can say that, on our assumptions, the SG fact grounds the supremacy of morality, which is to say that it grounds the conclusiveness of all-in moral reasons, including the reason it itself is to do what one ought morally to do rather than what one ought-X to do. On our assumptions, the SG fact would ground the conclusiveness of the moral reason that *it itself is* for anyone to do what she ought morally to do. We are citing the SG fact to ground the fact that the SG fact itself is a normatively conclusive reason. This is a bootstrapping response. The conclusiveness of a reason, if there are any conclusive reasons, would be grounded by some *further* fact about the fact that is the reason. It is not something that could be grounded in the very fact that is the reason.[8]

Suppose, for the sake of illustration, that the SG fact is the fact that the currency of morality will maximize the general welfare. This, we assume, is the fact that grounds the supremacy of morality, and it is also the reason one has to act morally. That is, the fact that the currency of morality will maximize the general welfare is the reason to act morally *and* this fact makes it the case that the reason *that it itself is* to act morally is conclusive. We are trying to lift ourselves by pulling on our boots.

There is, however, the alternative that we set aside before. Perhaps it is *not the SG fact*, but rather, it is *the fact that morality is the supreme standpoint*, that is the conclusive reason for Alice to do what she ought morally to do rather than what she ought-X to do. But the fact that morality is "supreme" is by stipulation the fact

[7] Killoren (2019) seems to think that such a regress would not be vicious. I disagree. If there is a regress of standpoints of the kind at issue, the supremacy of S is ungrounded. S's supremacy is not grounded unless R grounds its authority, and R cannot do this unless Q grounds its authority, and so on, without end.

[8] I am here responding to an objection raised by Tristram McPherson.

that what one all-in ought morally to do is what one ought *simpliciter* to do. And the "conclusiveness" of an all-in reason is by stipulation the fact that this is the all-in reason to do what one ought *simpliciter* to do. So, given these stipulations, the fact that morality is supreme entails that there is a conclusive all-in reason to do what one all-in ought morally to do to do. It entails that *there is* a conclusive reason for Alice to do what she ought morally to do but it is not itself that reason. Alice's question is, what is this reason? The response that morality is supreme tells her there is a reason without telling her what the reason is.

So, if we assume that the normative properties of a thing must be grounded by or explained by other facts about the thing, we can stipulate that the SG fact is whatever fact grounds or explains S's superiority. To ground S's superiority, the SG fact must ensure that agents have conclusive reason to do what they ought-S to do rather than what they ought-X to do. But at the same time, the SG fact must *be* this reason. The lack of a non-circular account of what grounds the conclusiveness of this reason is the lack of a non-circular account of the superiority of S, or of the fact that agents ought *simpliciter* to do what they ought-S to do. This is a problem.

I said before that, to challenge my argument, one might question my assumption that the normative properties of a thing must be grounded by, or explained by, other facts about the thing. Consider, then, the view that morality is supreme or overriding but that this fact is not grounded by any other fact about morality. There is no supremacy grounding property. There is just the "brute" fact that morality is overriding. There is no explanation for this and no need to explain it. Perhaps the property of normative supremacy is primitive, unanalyzable, and ungrounded. If this is correct – if morality is supreme, but there is no supremacy grounding property – my argument is in trouble.[9]

In response, I suggest that the general thesis that the normative properties of things are grounded in other facts about those things is philosophically very attractive.[10] And I can see no argument for making an exception of the property of normative supremacy. Moreover, the "ungrounded supremacy" position has at least two unattractive implications. It implies that there is no answer to the question of why morality is supreme, and no answer to the question of why self-interest is not supreme. It also implies that certain "why be moral" questions are unanswerable. Alice asks, "What is the conclusive reason for me to do what I am morally required to do?" The claim that morality is overriding does not answer the question, for it amounts simply to saying that there is a conclusive reason to do what one is morally required to do. And on the ungrounded supremacy view, morality can be overriding without this being grounded in any fact about morality that might provide the reason. So, it seems, on this view, Alice's question has no answer. The view I recommend says that, if morality is overriding, the supremacy grounding fact is the reason Alice is looking for. The problem is that there

[9] Ruth Chang posed this objection.

[10] This thesis is compatible with the claim that normative properties are unanalyzable.

seems to be no non-circular account of what grounds the conclusiveness of this reason. So there is no non-circular account of the superiority of morality.

Does my argument show that there is not a supreme standpoint, that there is nothing an agent ought *simpliciter* to do unless all relevant normative standpoints agree? What it does, I think, on the assumption that pluralism is correct, is to cast doubt on the idea that there is a way to explain or ground the supremacy of a normative standpoint – that there is a way to explain the putative fact that some standpoint is such that the verdicts it produces as to what one ought to do are verdicts as to what one ought *simpliciter* to do. So unless we think there can be brute normative facts that are not grounded or explained by other facts, we should be skeptical of the thesis that there is a supreme standpoint. We should incline to strong pluralism.

We should ask what property a standpoint S might have such that facts as to what a person ought-S to do would be facts about what the person ought *simpliciter* to do. I have called this the "supremacy property," but this is only to give it a name. We can articulate what the existence of a standpoint with this property would do for us, if pluralism is correct. We have a job description. But I doubt that we have any idea of what property could do this work. And, it seems to me, on the assumption that pluralism is correct, we have only a muddled idea of what could make a claim regarding an ought *simpliciter* be true – setting aside cases in which all relevant normative standpoints agree. I now want to suggest that the thesis that there are oughts *simpliciter* is a theoretical postulate that we do not need.

6 "Metaphysical" and "Evaluative" Propositions about *Ought*

We all have value commitments that give greater priority to some kinds of concern over others. Many of us rank moral concerns as more significant than our own self-interest except in cases in which morality would ask us to make a highly significant sacrifice. Many others rank self-interest as more important than moral considerations except perhaps in cases where morality asks little of us. Now, according to strong pluralism, there is not a supreme normative standpoint that would validate the ranking that a person's value commitments give to various types of consideration. Nothing in the nature of things validates this ranking. Still, a pluralist could and should take a person's value commitments to constitute a normative standpoint, the standpoint of her values. A person's value commitments have implications for what she ought to do relative-to-her-values and for the reasons she has relative-to-her-values.[1]

We can accordingly distinguish two families of propositions. One is a family of propositions, each of which presupposes there is a supreme standpoint. I will label these "metaphysical." An example is the proposition that Alice ought *simpliciter* to make the donation. Assuming pluralism, and given that Alice is in a conflict situation, this proposition presupposes that there is a supreme standpoint. The other is a family of "evaluative propositions" about reasons and *oughts* relative-to-one's-values. An example is the proposition that Alice ought to make the donation *relative-to-her-values*. There are corresponding families of beliefs. "Metaphysical beliefs" take metaphysical propositions as their content, and "evaluative beliefs" take evaluative propositions as their content.

Suppose I think that Gyges simply ought not to murder the king. My belief could be the metaphysical belief that Gyges ought *simpliciter* not to murder the king. Or it could be the evaluative belief that Gyges all-in ought not to do this relative-to-my-values. Both of these could be expressed in an appropriate context by saying "Gyges simply ought not to murder the king." According to strong pluralism, the metaphysical belief is not true. Hence, the charitable interpretation is that my belief is the evaluative one. There is no conflict between strong pluralism and the belief we might express by saying, "Gyges simply ought not to murder the king," if this belief is an evaluative belief, in which the *ought* is relative to values we share rather than the metaphysical belief that presupposes the existence of a supreme standpoint.

[1] Note that there can be a reason-relative-to-*my*-values for *you* to do something. Reasons-relative-

There are three reasons to accept the charitable interpretation of our intuitions in conflict cases. First, it is typical to make judgments and decisions about what to do relative to one's own evaluative standpoint (see Copp 2009). This, I think, is common ground between pluralists and unifiers. If one of Alice's eligible options is such that, relative to her own values, she ought to do that thing, she likely would decide to do it. And having made her decision, she might announce that this is what she simply ought to do. Second, it seems unlikely that people have pre-theoretical beliefs the truth of which depends on the existence of a supreme standpoint. Suppose Alice decides that she simply ought to make the donation. If we objected on the ground that there is no supreme standpoint to settle what a person ought *simpliciter* to do, she almost certainly would be taken aback. She would likely deny that the truth of her view depends on the existence of any such thing. Third, in expressing thoughts about what we ought to do, we often do not explicitly mention the standpoint relative to which we ought to do the thing. (Similarly, we typically do not mention that the weights of vegetables in the corner store are relative to the Earth's gravitational field.) If Alice says, "I ought to make the donation," the belief she expresses might be one in which the *ought* is relative to her own values. *Bare* "oughts" do not necessarily express propositions about oughts *simpliciter*.

Perhaps it will seem, however, that the evaluative propositions in question actually are propositions about what we ought *simpliciter* to do.[2] Suppose that Alice decides to upgrade her ticket rather than to make the donation. She presumably was led to her decision by her values, and from her perspective, her decision settles what to do. So, it might seem, her decision expresses her judgment about what she ought *simpliciter* to do. That is, given my earlier stipulation, her decision expresses her judgment about what she ought finally, period, and in the end to do, which settles what to do.[3]

If this were correct, I would need to reword my claims, but I believe it is a mistake. It is not, at any rate, how we think about evaluative judgments. First, we are capable of acting contrary to our evaluative judgments, and when we do, provided we are reasonably self-aware and humble, we need not think of ourselves as acting contrary to what we ought finally, period, and in the end to do. For example, Alice might conclude that, in light of her values, she ought to upgrade her ticket, but she could instead decide to make the donation. And, in doing this, she might deny that she is failing to do what she ought *simpliciter* to do. She might instead reconsider her values and decide to accord more importance to morality. Second, when other people think about Alice and her values, they would not ordinarily think that the standpoint of *her* values determines what she ought *simpliciter* to do. For they might not share her values. Third, it seems plain that an evaluative belief does not presuppose that there is a supreme normative standpoint, and there is no reason to think that the standpoint

---

to-one's-values therefore are not "internal" in Williams's sense (1981).

[2] This was suggested by Tristram McPherson and Andres Luco.

[3] See Baker (2017, sec 2.7) for the related idea that we might explain the special status of the ought *simpliciter* in psychological terms. See also Finlay 2013; Tiffany 2007; Hubin 2001: 465.

of Alice's values is normatively supreme. Yet it would have to be supreme in order for its verdict to settle what Alice ought to do, finally, period, and in the end.

Strong pluralism seems counter-intuitive to many people. But, I think, in many cases where it might seem that our beliefs conflict with strong pluralism, there is not actually any conflict. We rather have evaluative beliefs the truth of which does not depend on the existence of a supreme standpoint. We are conflating these beliefs with metaphysical beliefs that do presuppose this.

7 Objections to the strong pluralist view

A variety of objections have been raised against strong pluralism. In this section, I will address the three most important. I believe the objections lose much of their force when we bear in mind my distinction between evaluative beliefs and so-called metaphysical beliefs. As I said, I believe that the idea of an ought *simpliciter* is a theoretical idea that we do not need.

The first objection is that strong pluralism cannot make sense of the idea of practical reason. According to strong pluralism, practical reason is merely one of the competing normative standpoints, on a par with morality and self-interest. According to strong pluralism, then, one might object, a person in a conflict situation must simply decide *arbitrarily* what she will do (McPherson 2017). Her decision is arbitrary even if she decides to do what practical reason recommends. But this seems incoherent, for it is not *arbitrary* to decide to do what practical reason recommends. Indeed, one might think, it is plausible that the standpoint of practical reason is the supreme standpoint.

Let me first respond to the claim about arbitrariness. Strong pluralism does *not* imply that a person in a conflict situation must make an arbitrary decision (see Sagdahl 2016). Alice would not be deciding arbitrarily if she decided to do the thing that is most strongly supported by her values. Indeed, I think, Alice might be entirely rational in deciding what to do, assuming she takes into account the reasons of different kinds that bear on her decision, and assuming her decision accords with her values (see Copp 2007b). A person can act non-arbitrarily without being guided by a belief about what she ought *simpliciter* to do.

Next turn to the suggestion that the standpoint of practical reason is the supreme standpoint. Given familiar accounts of what practical reason consists in, I think this is highly implausible. Consider, first, the minimalist view according to which practical rationality is simply a matter of a kind of psychological coherence (see Broome 2013; Scanlon 1998). In this view, plausibly, practical reason *is* merely one normative standpoint, on a par with morality and self-interest. For there can be cases in which a person ought morally, or ought in her self-interest, to allow herself a kind of psychological incoherence.[1] Perhaps Alice can prevent an innocent person's begin murdered only by being irrationally incoherent. In such a case, it is implausible that Alice ought *simpliciter* to sustain her psychological coherence. So, in the minimalist model,

[1] Consider some of the examples given by Schelling (1960).

it is implausible that practical reason is the supreme standpoint. Consider, next, the view that practical rationality is a matter of acting in one's self-interest. In this view, practical reason would compete with morality in typical conflict situations. But it is completely implausible that, for example, the fact that it is in Gyges's self-interest to go ahead with his plot settles that this is what he ought *simpliciter* to do. So it is implausible, in this theory, that the rational standpoint is the supreme standpoint. In short, it seems, strong pluralism can plausibly deny that the standpoint of practical reason is the supreme standpoint.

The second objection is the objection from "nominal/notable comparisons."[2] A "nominal/notable situation" is a conflict situation in which an agent has two eligible options, one of which is "notable," in that it is supported by very strong reasons of some kind, and the other of which is "nominal," in that it is supported by extraordinarily weak reasons of another kind. Imagine that Gyges* has a minor itch that, weirdly, he can stop, at no risk to himself, only by murdering the king. Gyges* ought in his self-interest to murder the king, but morally he ought not to do this. It seems intuitively clear that Gyges* simply ought not to murder the king – his itch is too trivial to justify a murder. For strong pluralists, however, it is not the case that Gyges* ought *simpliciter* to act morally even though his conflicting self-interested reason is extraordinarily weak. According to strong pluralism, it is not the case, in *any* conflict situation, that a person ought *simpliciter* to act morally. This is difficult to accept.

A related objection is that, according to strong pluralism, the relative normative force of conflicting moral and self-interested considerations in conflict situations does not depend on how strong they are. The reasons of self-interest that favor Gyges's carrying out his plot in the original example are very strong, whereas, in the itch example, the reasons of self-interest that favor his carrying out his plot are very weak. According to strong pluralism, even if everything else is held the same, the difference in the strength of these reasons makes no difference to how Gyges is to act.

The best response to the nominal/notable objection is to treat it as a first-order objection grounded in our values. Strong pluralists can agree that only a person who was morally bankrupt would murder someone to relieve an itch. Furthermore, strong pluralists can agree that everyone except a morally bankrupt person would have values such that, relative to their values, Gyges ought not to murder the king. Plausibly, I think, our intuition that an agent in a nominal/notable situation *simply ought* to do the notable thing is an evaluative belief in which the *ought* is relative to values we share. The fact that we share *this* thought is no barrier to accepting strong pluralism.[3]

The third objection concerns the nature of deliberation and decision-making. If strong pluralism is true, one might think, we cannot make sense of deliberation and decision-making and we cannot make sense of the related practice of giving advice.

[2] The objection was first proposed by Chang (1997: 32–34). It has been widely discussed. For example, see Dorsey 2013; Sagdahl 2013, ch 7 and 2014.

[3] Sagdahl suggests a similar strategy for responding to the objection (2013: 156–157), and he discusses various other strategies (2013: ch 7 and 2014).

When we are deliberating, trying to decide what to do, it is as if we were giving advice to ourselves. And plausibly, in giving advice, we are expressing the thought that our advisee ought *simpliciter* to do something. Otherwise our advice would be insincere. If so, it appears that strong pluralism cannot make sense of advice-giving in conflict situations (see Thomson 2001: 46). It follows that it cannot make sense of deliberation. In deciding what to do, our aim is to do the right or best thing *period*, not merely the thing that is right or best in relation to our values. Evidence of this is that we think we might make a mistake in deciding to do what would in fact best serve our values. We are not aiming merely to serve our values. Yet according to strong pluralism, in conflict situations, there is no fact as to what would be best or right period, in the end, and *simpliciter*.[4]

Let me first respond to the claims about advice-giving. I do not agree that, in giving advice, we invariably express the thought that the advisee ought *simpliciter* to do something. Advice is relative to a standpoint even if we do not have this explicitly in mind. When Alice asks for advice, for example, it might be clear that she is asking what we would do in her situation. Or she might be asking how to most further her values, or what she ought morally to do. In each of these scenarios, we could give our advice without mentioning the relevant standpoint, but the fact that we do not mention a standpoint does not mean that our advice is not relative to one.

Next consider deliberation and decision-making. I think strong pluralism can account for how things seem to us in deliberation without attributing systematic error to us and without supposing that our thoughts presuppose the existence of a supreme standpoint. For, I maintain, it is not the case that, in deliberating, we aim to do the right or best thing *simpliciter*. Further, I think, strong pluralism can account for the thought we might have, in deliberating, that our decision might be mistaken even if we decide to do what we correctly decide would best serve our values.

Strong pluralists should say that, in deliberating, we normally aim to do the right or best thing *relative to our values*, even if we do not have this explicitly in mind. Or, rather, they should say that this is how our deliberation is normally *best understood*. The fact that some action is best relative to our values would normally be in the background of our thinking, in the sense explained by Pettit and Smith.[5] There is the general point that the relativization of reasons and *oughts* to normative standpoints is not typically explicit in our thinking. But further, as Pettit and Smith would suggest, when we decide what to do, we do not normally cite the fact that we value something as a reason. If I decide to go to a jazz club, I might cite, as my reason, the fact that *this is a good place to hear jazz*. I would not normally cite, as my reason, the fact that *I value listening to jazz and by going to this club I can further this value*. My values are in the background, shaping my deliberation. My deliberation normally mentions things

[4] Stephen Darwall and Gary Watson helped me to understand this objection.

[5] Pettit and Smith (1990) were discussing the backgrounding of desire, but the issues are the same. See esp. pp. 574–577.

I value without citing the fact that these are things that I value. My reason to go to the club is relative to my values even if this fact is not something I take into account in my deliberation. This illustrates a role played by our values in our deliberation. It illustrates my claim that our aim in deliberation is normally best understood as the aim to do the right or best thing relative to our values – even though the relativity to our values is normally not explicit in our thinking and even though facts about our values are normally in the background.

When we deliberate, we want to get things right, and we sometimes worry we might be making a mistake. Strong pluralism should say we might make many different kinds of mistakes. Of course, if we have the explicit aim to further our values, we might worry about being mistaken about how to achieve this aim. We can fail to understand what we truly value. We can fail to appreciate how the circumstances we face would best be dealt with, given our values. We can do something that is best relative to our current values but that will not stand the test of time, as our values mature. Similarly, from the perspective of our current values, we might have been mistaken in some of our past decisions. And if we have an explicit aim other than that of serving our values, we might worry about corresponding kinds of mistakes. Suppose we have the explicit aim of doing what is morally right. In deliberating about this, we will of course be guided by our own moral values, but our *goal* would be to do what is in fact morally right, rather than what our own moral values require. Alice might try to decide whether it would be morally best to donate her money to Oxfam. She might worry whether making the donation would be fair to herself. She might think she is morally permitted to use the money to further her own welfare. She would try to figure out, by her own best lights, what morality requires. This kind of reasoning is not a problem for strong pluralism.

When we try to decide what to do, we do not engage in an effort to discover what would be recommended by some supreme standpoint. We rather consider our circumstances, and the facts of the case, in light of our values. If we think we are making the wrong decision, we might try to be more careful in our deliberation. We can step back from the case at hand, if a decision is not pressing, and reconsider our values or our priorities. Moral argument can lead us to change our moral values. We can also change our values in light of a better understanding of what our current values commit us to. We can be impressed by the example given us by a person we admire, and decide to emulate her. All of this makes sense on the strong pluralist view, and I think this vindicates at least much of the phenomenology of deliberation. There is no "Archimedean point" and, ordinarily, we do not look for one.

To summarize, it seems to me that strong pluralism can handle the three main objections. The objections lose much of their plausibility when we bear in mind the distinction between evaluative propositions and propositions the truth of which depends on the existence of a supreme standpoint.

8 Conclusion

I began by arguing for pluralism, and then, taking pluralism as a premise, I argued for strong pluralism, for the thesis that there is not a supreme standpoint. Strong pluralism can allow that there are facts about what one ought *simpliciter* to do in cases where all relevant normative standpoints agree. The interesting question, however, is whether there are oughts *simpliciter* in conflict situations, and strong pluralism implies there are not, since there is no supreme standpoint. The idea that there is such a standpoint is, I think, an elusive theoretical postulate. Yet strong pluralism does not have the counter-intuitive implications that it might seem to have. In our ordinary lives, when we are deciding what to do, our concern is to advance the things that we value – understood *de re* rather than *de dicto* – and also, I think, to meet our needs. We decide from the perspective of our values. Doing so does not presuppose that there is a supreme normative standpoint.

At the outset, I said my goal would be to explicate certain thoughts that Alice and Gyges might have in deciding what to do in the conflict situations they face. These are thoughts they might express by saying something like, "This is what I ought simply to do." What is the content of such thoughts? If strong pluralism is true, it would be uncharitable to interpret these thoughts as beliefs about what they ought *simpliciter* to do. Plausibly, instead, their thoughts, with their contents fully spelled out, would be judgments as to which of their eligible options is best supported by their values. If strong pluralism is true, it would be uncharitable to take their thoughts to presuppose the existence of a supreme standpoint.

Acknowledgments

A version of this chapter was presented to the Departments of Philosophy at Nanyang Technological University, Singapore, and at National Chung Cheng University, Taiwan. I would like to thank everyone who contributed to discussion on these occasions for their helpful comments. I am grateful in addition to Ruth Chang, Stephen Darwall, Andrew Forcehimes, Andres Carlos Luco, Tristram McPherson, David Plunkett, Mathias Sagdahl, Kurt Sylvan, Jon Tresan, Peter Shiu-Hwa Tsu, and Gary Watson for their helpful comments and suggestions.

References

Baier, Kurt. 1958. *The Moral Point of View*. Ithaca: Cornell University Press.

Baker, Derek. 2017. "Skepticism About 'Ought' *Simpliciter*." In Russ Shafer-Landau, ed., *Oxford Studies in Meta-Ethics*, 12. Oxford: Oxford University Press.

Broome, John. 2013. *Rationality Through Reasoning*. Oxford: Wiley Blackwell.

Chang, Ruth. 1997. "Introduction", in R. Chang, ed., *Incommensurability, Incomparability, and Practical Reason*, 1–34. Cambridge: Harvard University Press.

———. 2002. "The Possibility of Parity." *Ethics*, 112: 659–688.

Copp, David. 1995. *Morality, Normativity, and Society*. New York: Oxford University Press.

———. 2007a. *Morality in a Natural World*. Cambridge: Cambridge University Press.

———. 2007b. "The Normativity of Self-Grounded Reason." In D. Copp, ed., *Morality in a Natural World*, 309–354. Cambridge: Cambridge University Press.

———. 2007c. "The Ring of Gyges: Overridingness and the Unity of Reason." In D. Copp, ed., *Morality in a Natural World*, 284–308. Cambridge: Cambridge University Press.

———. 2009. "Toward a Pluralist and Teleological Theory of Normativity." *Philosophical Issues*, 19: 21–37.

———. 2014. "Indirect Epistemic Teleology Explained and Defended." In Abrol Fairweather and Owen Flanagan, eds., *Naturalizing Epistemic Virtue*, 70–91. Cambridge: Cambridge University Press.

———. 2015. "Explaining Normativity." *Proceedings and Addresses of the APA*, 89: 48–73.

Dancy, Jonathan. 2006. "Nonnaturalism." In David Copp, ed., *The Oxford Handbook of Ethical Theory*, 122–145. New York: Oxford University Press.

Darwall, Stephen. 2006. *The Second Person Standpoint*. Cambridge: Harvard University Press.

Dorsey, Dale. 2013. "Two Dualisms of Practical Reason." In Russ Shafer-Landau, ed., *Oxford Studies in Meta-Ethics*, vol. 8, 114–139. Oxford: Oxford University Press.

Finlay, Stephen. 2013. *Confusion of Tongues: A Theory of Normative Language*. New York: Oxford University Press.

Foot, Philippa. 1978. "Morality as a System of Hypothetical Imperatives." In *Virtues and Vices*, 157–173. Oakland: University of California Press.

Gauthier, David. 1986. *Morals by Agreement*. Oxford: Oxford University Press.

Hubin, Donald. 2001. "The Groundless Normativity of Instrumental Reason." *The Journal of Philosophy,* 98: 445–468.

Killoren, David. 2019. "Infinitism About Cross-Domain Conflict." In Russ Shafer-Landau, ed., *Oxford Studies in Meta-Ethics*, vol. 14, 144–167. Oxford: Oxford University Press.

Korsgaard, Christine. 1986. "Skepticism About Practical Reason." *The Journal of Philosophy*, 83: 5–25.

———. 1996. *The Sources of Normativity.* Cambridge: Cambridge University Press.

McCleod, Owen. 2001. "Just Plain Ought." *The Journal of Ethics*, 5: 269–291.

McDowell, John. 1978. "Are Moral Requirements Hypothetical Imperatives?" *Aristotelian Society, Supplementary Volume*, 52: 13–42.

McPherson, Tristram. 2017. "Authoritatively Normative Concepts." In Russ Shafer-Landau, ed., *Oxford Studies in Meta-Ethics*, 12. Oxford: Oxford University Press.

Parfit, Derek. 1984. *Reasons and Persons.* Oxford: Oxford University Press.

———. 2011. *On What Matters*, 2 vols. Oxford: Oxford University Press.

Pettit, Philip, and Michael Smith. 1990. "Backgrounding Desire." *The Philosophical Review*, 99: 565–592.

Sagdahl, Mathias Slattholm. 2013. *The Prospects of Normative Pluralism.* PhD dissertation, University of Oslo, Oslo.

———. 2014. "The Argument from Nominal-Notable Comparisons, 'Ought All Things Considered,' and Normative Pluralism." *The Journal of Ethics*, 18: 405–425.

———. 2016. "Enkratic Reasoning and Incommensurability of Reasons." *Journal of Value Inquiry,* 50: 111–127.

Scanlon, T. M. 1998. *What We Owe to Each Other.* Cambridge: Harvard University Press.

———. 2014. *Being Realistic About Reasons.* Oxford: Oxford University Press.

Schelling, Thomas. 1960. *The Strategy of Conflict.* Cambridge: Harvard University Press.

Schroeder, Mark. 2007. *Slaves of the Passions.* Oxford: Oxford University Press.

Sidgwick, Henry. 1981. *The Methods of Ethics*, 7th edition. (First published, 1907; first edition, 1874.) Indianapolis: Hackett Publishing Company.

Skorupski, John. 2010. *The Domain of Reasons.* Oxford: Oxford University Press.

Smith, Michael. 1994. *The Moral Problem.* Oxford: Blackwell.

Tiffany, Evan. 2007. "Deflationary Normative Pluralism." *Canadian Journal of Philosophy*, 37 Supplement, 33: 231–262.

Thomson, Judith Jarvis. 2001. *Goodness and Advice.* Princeton: Princeton University Press.

Wedgwood, Ralph. 2007. *The Nature of Normativity.* Oxford: Oxford University Press.

Williams, Bernard. 1981. "Internal and External Reasons." In *Moral Luck*, 101–113. Cambridge: Cambridge University Press.

Wolf, Susan. 1982. "Moral Saints." *The Journal of Philosophy*, 79: 419–439.

29. There is no moral *ought* and no prudential *ought*

Elizabeth Harman

1 Introduction

It is natural to think that there are a number of different *oughts*. There is a moral *ought*, there is a prudential *ought*, and so on. Furthermore, it is natural to think that each *ought* is such that one ought to do the best thing one could do, where the sense of *best* at issue varies with the kind of ought it is.[1] Thus, it is natural to think: morally, a person ought to do the morally best thing she could do, and prudentially, a person ought to do the thing that is best for her. One might also express these thoughts by saying: morality recommends that one do the morally best thing one could do, and prudence recommends that one do the thing that is best for oneself.

These natural thoughts suggest the further thought that the moral *ought* and the prudential *ought* often conflict, and thus that often, morally, one ought to do something although, prudentially, one ought to refrain from doing it.

While these thoughts are natural, and they express a commonly held view, I will argue that these thoughts are wrong. My modest aim is to show that there is an alternative view to the view I have described. My more ambitious aim is to show that my alternative view is correct. Once the two views are contrasted, I think it will be clear that although the commonly held view is indeed common, it is not supported by or warranted by ordinary moral thinking or ordinarily recognized moral phenomena, and we do better at capturing moral reality – and normative reality more broadly – by jettisoning the common view in favor of the alternative view I outline.

The alternative view denies all the natural thoughts I have described. It holds that there is no distinctively moral *ought*, though there are some *ought* facts that are

[1] For example, in *Opting for the Best* (New York: Oxford University Press, 2019), Douglas Portmore explicitly develops a view along these lines, arguing that – for maximal (complete) options that one has at a time – morally, one ought to take an option just in case it is one's morally best option, and prudentially, one ought to take an option just in case it is one's prudentially best option, and Portmore argues that these claims hold, no matter what makes an option morally best (thus, no matter whether morality involves any agent-relative constraints or permissions).

In another example, in his work on the supererogatory, Paul McNamara uses "ought" to pick out an agent's morally best option. (See his "Supererogation: Inside and Out," *Oxford Studies in Normative Ethics* 1 (2011): 202–235, and other work.)

In a third example, Stephen Finlay argues that what one "morally ought to do" is the morally best thing one could do, in his "Too Much Morality," in *Morality and Self-Interest*, ed. Paul Bloomfield (New York: Oxford University Press, 2007), 136–154.

Work on the semantics of the word "ought" often assumes that "ought" picks out an agent's best option; note that my target is the more specific claim that there is a moral *ought* that picks out an agent's *morally best* option.

Note that the idea that, morally, an agent ought to take her morally best option does not commit

distinctively moral. Similarly, there is no distinctively prudential *ought*, though there are some *ought* facts that are distinctively prudential. Finally, the alternative view holds that distinctively moral *ought* claims never conflict with distinctively prudential *ought* claims: it is never the case that, morally, one ought to do something, while prudentially, one ought to refrain.

one to consequentialism: the morally best *option* is simply the morally best *thing to do*, and it need not be the option that has the best consequences. For example, one could hold that one's morally best option is always to keep one's promises, even in cases in which breaking a promise would have better consequences.

2 The common view: there is a moral *ought* and a prudential *ought*

Consider the following view:
The Common View:

> (a) There is a moral *ought* such that: Morally, an agent ought to do something, just in case morality recommends that she do it. *and* Morally, an agent ought to do something, just in case it is the morally best thing she could do. *and* Morality recommends that an agent do something just in case it is the morally best thing she could do.
>
> (b) There is a prudential *ought* such that: Prudentially, an agent ought to do something, just in case prudence recommends that she do it. *and* Prudentially, an agent ought to do something, just in case doing it would be best for the agent. *and* Prudence recommends that an agent do something just in case doing it would be best for the agent.
>
> (c) These two *oughts* often conflict: it is often the case that morally, one ought to do something, while prudentially, one ought to refrain from doing it.

I will argue that the Common View can be (and should be) rejected. I will present an alternative picture. On the alternative picture, it is indeed sometimes true that morally, one ought to do something, and it is sometimes true that prudentially, one ought to do something, but these truths do not involve a distinct moral *ought* nor a distinct prudential *ought*. The alternative picture embraces the first biconditional under (a), embracing the tight connection between what, morally, one ought to do and what morality recommends. And the alternative picture embraces the first biconditional under (b), embracing the tight connection between what, prudentially, one ought to do and what prudence recommends. But the alternative picture rejects the other two conditionals under (a) and the other two conditionals under (b): on the alternative picture, the asserted tight connections with the morally best thing the agent could do, and with what would be best for the agent, do not hold.

The plan for the rest of chapter is as follows. In sections 3, 4, and 5, I discuss and argue against the Common View, as it concerns morality. Sections 6 and 7 address objections. In section 8, I present the part of the alternative view that concerns morality.

In section 9, I argue against the Common View, as it concerns prudence. Section 10 presents the remaining part of the alternative view, concerning prudence. Section 11 discusses some objections. Section 12 summarizes the chapter's conclusions. And, finally, section 13 explains the broader significance of these conclusions.

3 Moral *ought* statements do not just state moral requirements

In this section, I will explain why people may have been drawn to the Common View. To see the appeal of the Common View, let's contrast it with another view. The Common View includes the following claim:

> (*) Morally, an agent ought to do something, just in case it is the morally best thing she could do.

The contrasting view is:

> The Naïve View: Morally, an agent ought to do something, just in case she is morally required to do it.

I call this view "naïve" because I don't think any non-consequentialist holds it. (It would, however, be endorsed by a maximizing consequentialist.)

These two views agree about cases involving morally required options. For example, consider this case:

> Aaron promised Bill that he would go to Bill's poetry reading. Aaron does not enjoy poetry, but Bill is his friend, and it will mean a lot to Bill to have Aaron there. On the day of the reading, Aaron is invited to join another friend at a Lakers game. Aaron loves the Lakers, though of course there are other games he could attend this season. Aaron tells all this to Carl. Carl says, "Sorry! I know it's tempting to bail on Bill, but morally, you ought to go to the reading."

In this case, it is true that, morally, Aaron ought to go to the reading.[1] Both the Common View and the Naïve View accommodate this truth, because going to the reading is Aaron's best option, and it is also a morally required option. Similarly, consider this case:

> Donna is an amateur tennis player who maintains a fierce and genuinely hostile rivalry with Ellen, another tennis player. Donna is doing a pre-

[1] Note that Carl could also express this truth – the truth that, morally, Aaron ought to go to the

> liminary review of applications for a job at her company, after which a committee will make a decision among the strongest candidates. Donna sees that Ellen's application is very strong. Donna hates the idea of working with Ellen and knows that she could eliminate Ellen's application from consideration. Donna tells all this to Fiona. Fiona says, "Sorry! I know it's tempting to reject Ellen at this point, but if her application is strong, then morally, you ought to keep her under consideration along with the other strong applicants."

In this case, it is true that, morally, Donna ought to keep Ellen under consideration.[2] Both the Common View and the Naïve View accommodate this truth, because keeping Ellen under consideration is both Donna's best option and it is morally required. The same will hold for any case involving a morally required option. Whenever an agent has a morally required option, then that option is also her morally best option, and so both the Common View and the Naïve View will imply, correctly, that morally, she ought to take that option.

To see why one would reject the Naïve View in favor of the Common View, let's turn to considering some cases involving supererogatory actions. In ordinary life, we often say to each other, truly, "You *ought to* do it, but you don't *have to* do it," where by "you don't have to do it," we mean that it isn't morally required, and in saying "you ought to do it," we are making a *moral* claim. For example, consider the following case:

> Georgia's elderly neighbor Harriet is recovering from surgery. Georgia knows that, given her friendship with Harriet, she is morally required to visit her at least a few times a week, and Georgia has already visited Harriet several times this week. Today Georgia comes home from work and is a bit tired but sees Harriet's light on across the street. Georgia could stop by to see Harriet for ten minutes. It would lift Harriet's spirits and wouldn't cost Georgia very much; she'd just start cooking her dinner ten minutes later. The morally best thing Georgia could do at this moment is to go spend ten minutes with Harriet.

In this case, it is natural to think that Georgia is not morally required to spend ten minutes with Harriet, but that Georgia *ought* to spend ten minutes with Harriet. It is natural to think that, morally, Georgia ought to spend ten minutes with Harriet. It seems that *morality recommends* that Georgia spend ten minutes with Harriet. This thus seems to be a case in which morality offers a *recommendation* that is not a *requirement*. The Naïve View does not allow that moral *ought* statements can do this. According to the Naïve View, morality recommends all and only those actions that are

---

reading – by simply saying, "You ought to go to the reading."

[2] Note that Fiona could also express this truth – the truth that, morally, Donna ought to keep Ellen under consideration – by simply saying, "You ought to keep Ellen under consideration."

morally required. By contrast, the Common View can accommodate all of these claims: spending ten minutes with Harriet is Georgia's morally best option at the moment, and so, morally, she ought to take it, according to the Common View.

Consider another case of a supererogatory action, which also illustrates this point:

> It is November 2016. Isaac is a college student who is worried after Donald Trump's election to the U.S. presidency. He donates his limited spare money to Planned Parenthood and the ACLU. He hears that some Muslim students are afraid of hate crimes and that some students are volunteering to walk them home from prayer at the local Mosque. It's raining and Isaac is tired. But this is a concrete way he could help. "I ought to go volunteer. I don't have to, but I ought to do it," he thinks.

Isaac is making a moral claim, and it is a moral claim that might well be true. The Naïve View cannot accommodate the truth of Isaac's *ought* claim, because volunteering is not morally required. But the Common View can accommodate the truth of his claim: morally, Isaac ought to volunteer.

The cases of Georgia and Isaac show that the Naïve View is false. Sometimes, morally, one ought to do something, although it is not morally required. These cases support the Common View instead.[3]

The Common View is indeed common; I hypothesize that philosophers have embraced the Common View by overgeneralizing from cases like Georgia and Isaac.

Let's take stock. The Common View includes these two claims:

> (*) Morally, an agent ought to do something, just in case it is the morally best thing she could do.
>
> (**) Morality recommends that an agent do something just in case it is the morally best thing she could do.

As we've seen so far, true instances of these claims are given by morally required actions, such as Aaron's going to the reading and Donna's keeping Ellen under consideration, and by some supererogatory actions, such as Georgia's spending ten minutes with Harriet and Isaac's volunteering to walk the Muslim students home. In the next section, I will argue that some instances of claims (*) and (**) are false, and thus that the Common View must be rejected.

[3] At this point in the dialectic, while discussing why we might be drawn to the Common View and should reject the Naïve View, I am simply taking these *ought* claims about Georgia and Isaac to be *true* in these cases. But I should mention that my considered view is *not* that these claims are obviously true given the details of the cases as they've been described. Rather, my considered view is that there are possible versions of these cases (as described) in which these normative claims are true. (That would be enough to imply that the Naïve View is false.)

4 Sometimes (often), morality does not recommend that one do the morally best thing one could do

Consider this case involving a supererogatory option:

> James is an accountant who paints paintings as a hobby. He has always wanted to have a show in a professional gallery but has not ever had one. The local gallery offers a competition: artists can submit one piece, and if their piece is selected, they get a show for one week in the gallery. The gallery is a popular destination for other gallery-owners and art buyers, so a one-week show there might lead to further successes as a professional artist. Getting the show would be the accomplishment of a long-held dream for James and might lead to further meaningful accomplishments. James works very hard on his piece to submit. On the day of the deadline, James is getting in his car to drive to the gallery, with the deadline in half an hour. James's neighbor Kenny, who is seven years old, wanders over and asks if James will play Parcheesi with him. James and Kenny do play Parcheesi sometimes. Kenny's mom has been sick, and Kenny is often sad and lonely, so it means something to Kenny when James plays with him. James is about to leave town for a work trip, so this would be his last chance to play with Kenny for a while. James has two options.
>
> (i) James could play Parcheesi with Kenny, missing his chance to submit to the art show.
>
> (ii) James could tell Kenny he can't do it today and go on to the art gallery.

In this case, what is James's morally best option? Well, staying and playing with Kenny would be a morally good thing to do. But going to the art gallery, while there is a lot to be said in its favor, is not a *morally* good thing to do. And whenever one option is morally good but an alternative is not morally good, then the first option is morally better. Because these are James's only two options, it follows that playing with Kenny is James's morally best option.[1] But it would be a big mistake for James

[1] An objector might say that James's morally best option is not playing with Kenny but rather doing something else entirely, like driving to the nearest hospital to offer up one of his kidneys to a

to stay to play with Kenny. This is a case in which taking his morally best option would involve James making a big mistake.

Now, suppose that morality always recommends that one take one's morally best option, as claim (**) says. If so, then this is what morality would say to James: "James, play with the neighbor boy." or "James, we recommend that you stay to play with Kenny, though you don't have to do that." Surely morality does not say this. Only a jerk would give James this advice. The advice that morality gives is not advice that only a jerk would give. Thus, this case shows that claim (**) is false. Furthermore, what morality recommends is exactly what, morally, an agent ought to do. Thus, the case also shows that claim (*) is false. It is not true that, morally, James ought to stay to play with Kenny. Thus, it is not always true that, morally, one ought to take one's morally best option; claim (*) is false.

Consider the following case, which also involves a supererogatory option:

> Laura is a distinguished professor who has had a long day at work, at the end of which she has a meeting with Mark, a young man in his mid-twenties who is seeking her advice on pursing a career in her field. Toward the end of the meeting, Laura mentions that she has a six-month-old baby. Mark says "Oh, so you're not really writing right now!" Laura can tell that he is attempting to be friendly and to have a moment of human connection, but she's offended and annoyed by his remark. His remark is sexist and, in her case, inaccurate. Laura has several options at this moment.
>
> (i) She could kindly take the time to explain to Mark why what he said is problematic, thus perhaps saving him from making future comments that might hurt him professionally and thus potentially saving some women from being the recipients of such comments, while offering the explanation in such a nice way that she does not make Mark feel too bad.
>
> (ii) She could proceed as though she is not at all offended and end the meeting on a friendly note.
>
> (iii) She could reveal her annoyance, ending the meeting politely but not warmly, and get back to her work.
>
> (iv) She could express her annoyance, pointedly and sharply, not sparing Mark's feelings.

Those are Laura's four options. Laura is tired after a long day and does not feel like spending her limited mental energy on educating yet another naïve young man. She takes option (iii), ending the meeting politely but not warmly.

In this case, all four of Laura's options are morally permissible. But option (i) is her morally best option: it would be a morally good and kind thing for her to do, though

---

stranger. My response is that whether an option is a person's morally best option depends on how it

it is not morally required. Nevertheless, morality does not *recommend* that Laura take option (i). If morality did make this recommendation, morality would be saying to Laura, "Laura, educate the young man in your office." or "Laura, although you are tired, and you have to deal with sexist stuff like this all the time, we recommend that you take the time to kindly explain to this man why what he said is problematic, and do it in a way that spares his feelings as much as possible." Only a jerk would say this to Laura. The recommendations of morality are not recommendations that only a jerk would make. So this case shows that claim (**) is false: morality does not always recommend that one do the morally best thing one could do. Furthermore, what morality recommends is exactly what, morally, an agent ought to do. Thus, it is not the case that, morally, Laura ought to take option (i). So this shows that claim (*) is false: it is not always true that, morally, an agent ought to take her morally best option.

In this section, I've argued for the strong conclusion that the Common View is false. But we can also take this section as arguing for the weaker conclusion that a rejection of the Common View has some plausibility. If I haven't convinced you of my claims about the cases of James and Laura, I hope to have begun to show what an alternative view would look like and why it has some plausibility. I will elaborate on that alternative view in section 8.

---

compares to its alternatives: the common view holds that, morally, one ought to do something just in case it is morally better than any *alternatives to it*. It's true that donating a kidney is morally better than playing Parcheesi with Kenny, but donating isn't really an alternative to playing with Kenny: James could simultaneously play with Kenny and make a phone call to the hospital to arrange a kidney donation.

5 Sometimes morality's recommendation is to do a morally good thing that is not the agent's morally best option

In section 4, I argued that certain cases of the following kind provide counterexamples to the Common View: cases in which an agent has a morally good option, but morality does not recommend that the agent do something morally good. James could play with Kenny rather than submitting to the art contest, and Laura could kindly explain the sexist remark rather than end her meeting abruptly; these are morally good options available to these agents. Yet in these cases, there is no morally good option that is recommended by morality. In these cases, there is no morally good option such that, morally, the agent ought to take that option.

In this section, I will argue that another kind of case also provides a counterexample to the Common View: cases in which morality does recommend that the agent do something morally good but does not recommend her morally best option. In these cases, there is a morally good option such that, morally, the agent ought to take that option. Nevertheless, these cases provide counterexamples to the Common View.

Consider the following case:

> Nicole is a waitress at a restaurant where some of her regular customers are Deaf and communicate with each other in sign language. She wishes she knew sign language so that she could communicate with them in their primary language, in order to be friendly and welcoming on a regular basis. She looks into a class in American Sign Language, but the class would be expensive and the time it would take would cause her to lose valuable sleep. Instead, Nicole could study some YouTube videos that teach basic ASL signs. It would take some work, but she could learn enough to be able to greet her Deaf customers in ASL. Nicole thinks to herself, "I ought to study these YouTube videos, though I don't have to."

Nicole's claim is a moral claim, and it may well be true. Let's focus on a version of this case in which Nicole's claim is true. Morally, Nicole ought to watch the YouTube videos. This is a morally good thing that she could do and, morally, she ought to do it.

But watching the YouTube videos is not Nicole's morally best option. The morally best thing she could do in this case is to take the ASL class. Taking the ASL class would be an even nicer thing for Nicole to do for her Deaf customers. But even though this is the morally best thing she could do in this case, it is simply not true that, morally, she ought to take the ASL class. Furthermore, it is not true that morality recommends that Nicole take the ASL class. Instead, morality recommends that Nicole study the YouTube videos.

This is a case in which Nicole has a morally good option available – watching the YouTube videos – which is recommended by morality and which is such that, morally, she ought to take it. Nevertheless, this case shows that claims (*) and (**) are false, because Nicole's morally best option – taking the ASL class – is not such that, morally, she ought to take it and is not such that morality recommends that she take it.[1] Thus, this case shows that the Common View is false.[2]

If you're not convinced by what I say about Nicole's case, we can take her case as establishing the weaker claim that there is some plausibility to a denial of the Common View;[3] I elaborate on an alternative view in section 8.

---

[1] An objector might say that I have misunderstood what it is for an action to be the morally best thing an agent could do: what is it for an action to be morally best *just is* that the action is something that, morally, the agent ought to do. This objector agrees with my claims about whether, morally, an agent ought to do something but disagrees with my claims about which actions are morally best. My reply to this objection is that it divorces whether an action is morally good and morally better from whether an action is praiseworthy or how praiseworthy it is, whereas in fact there is a close connection between these things. For example, it is not the case that, morally, Nicole ought to take the ASL class, but she would be praiseworthy for doing so; indeed, she'd be more praiseworthy than she would be for studying the YouTube videos. Taking the ASL class is a morally better thing to do than studying the YouTube videos, as revealed by the difference in the praiseworthiness of the two actions.

[2] The existence of these "second-best" supererogatory actions, such as Nicole's watching the YouTube videos, is underappreciated in discussions of the supererogatory. For example, in Terry Horgan and Mark Timmons's, "Untying a Knot from the Inside Out: Reflections on the 'Paradox' of Supererogation," *Social Philosophy and Policy* 27 (2010): 29–63, they gloss the paradox of supererogation as arising because supererogatory actions are "morally best" (29). Similarly, Jamie Dreier's, "Why Ethical Satisficing Makes Sense and Rational Satisficing Doesn't," in *Satisficing and Maximizing*, ed. Michael Byron (Cambridge: Cambridge University Press, 2004), 131–145, proposes that supererogatory actions are those that are required from the perspective of beneficence, and Douglas Portmore's, "Are Moral Reasons Morally Overriding?" *Ethical Theory and Moral Practice* 11 (2008): 369–388, says "for there is a sense in which supererogatory acts are acts that agents morally ought to perform" (379). Paul McNamara's "Supererogation: Inside and Out: Toward an Adequate Scheme for Common Sense Morality" points out that sometimes a supererogatory action is not an agent's morally best option, that such an action can nevertheless be praiseworthy, and that this phenomenon is underappreciated.

[3] Consider the following objection: all that's going on in Nicole's case is that, given the expense and time-consumingness of the class, that option has already been ruled out, so it's true that, morally, Nicole ought to study the YouTube videos, simply because we are only evaluating a *proper subset* of her options, and it is the morally best of those options. The problem with this objection is that, if it were correct, then it would *also* be true that, morally, Nicole ought to take the class; this would be true relative to her whole set of options, but it is not true. (The objector is certainly right, of course, that sometimes it is

true that, morally, someone should do something, where this is a truth about doing that as compared to a subset of her options. For example, it is true that, morally, a murderer should act quickly (rather than painfully slowly). But in such cases, it is also true that, morally, the murderer should not kill anyone at all.) For a contrastive view of *ought*, see Stephen Finlay and Justin Snedegar's, "One Ought Too Many," *Philosophy and Phenomenological Research* 89 (1) (2014): 102–124. For discussion of the gentle murderer, see Frank Jackson, "On the Semantics and Logic of Obligation," *Mind* 94 (1985): 177–195.

6 Objection: the moral perspective ignores self-regarding reasons

Someone might object to the claims I've made about the cases of James, Laura, and Nicole as follows:

> Whether, morally, one ought to do something is of course different from whether one ought to do it, all things considered. It is not true that Laura or James, in their situations, ought to engage in helping behaviors, all things considered. But morally, that is what they ought to do, because that is the morally best thing that each can do. When we ask what an agent ought to do, morally, or (equivalently) what is the morally best thing she could do, or (equivalently) what morality recommends that she do, we must remember that morality takes a certain perspective. Consider the fact that James would be making a big sacrifice in staying to play with Kenny. This fact counts against staying, when we consider what James ought to do, all things considered. But it simply doesn't count that way when we are asking what the morally best thing James could do is; thus, it simply doesn't count that way when we are asking what James ought to do, morally.

This objector holds that the Common View is correct and that my claims simply misunderstand the nature of the moral perspective, from which morality's recommendations are issued. The objector makes the following claim:

> An agent's self-regarding (or self-interested) reasons against acting in certain ways simply don't count against acting in those ways from the moral perspective. That is, that an action would be burdensome to an agent simply doesn't count against doing it, from the perspective of morality.

But this claim is obviously false. Morality does see the force of self-regarding reasons against action.[1] Morality takes self-regarding reasons into account in failing to require

[1] We can distinguish reasons that are morally relevant from moral reasons. Self-regarding reasons are morally relevant, but they are (at least typically) not moral reasons: they do not tell in favor of an action's being morally good, morally bad, morally required, or morally wrong (thus, they are not moral reasons), though they may tell *against* an action's having one of these properties (thus, they are morally relevant). Douglas Portmore's "Are Moral Reasons Morally Overriding?" points out that we

certain actions, and in failing to prohibit certain actions, because of burdens that agents would otherwise be morally required to endure. For example, one is typically morally required to keep one's promises, but if keeping a promise would prove unexpectedly burdensome, breaking the promise is often morally permissible. If morality recommended to Laura and James that they perform the helping actions available to them, or if morality recommended to Nicole that she take the ASL class, then morality would be ignoring some considerations to which morality is otherwise sensitive.

should distinguish reasons that are morally relevant from moral reasons.

7 Objection: something can be good advice although only a jerk would offer it

In section 4, I argued that James and Laura both have supererogatory options that are the morally best things they could do but that it is false that, morally, they ought to take these options. I appealed to the fact that only a jerk would say to either James or Laura, "I recommend that you do this, despite the cost to you of doing it." An objector might respond as follows:

> Sometimes it is not okay to offer certain advice to someone, even though it's good advice. Sometimes only a jerk would urge someone to do something that she is morally required to do. Sometimes only a jerk would urge someone to do something that, morally, she ought to do.

I will offer two responses to this objection.

First, I want to clarify my attitude to the argument in section 4. In that section, I do indeed argue *from* the claim that only a jerk would say "I recommend that you do this" to the claim that it's false that morally, the agent ought to take that option. I do take that consideration to support that conclusion. But I also think that it is independently plausible that it is false that, morally, James ought to stay to play with his neighbor, and it is independently plausible that it is false that, morally, Laura ought to kindly explain to the young man why his comment is sexist, sparing his feelings as much as possible.[1] So, while I do argue for these moral claims in section 4 – and then go on to use these moral claims to argue that the Common View is false – ultimately I don't want all the argumentative weight of that section to fall on my claim about what follows from the fact that only a jerk would recommend these actions; I think my moral claims about these cases are independently plausible, and I'm happy to rely on them directly, in arguing against the Common View.

Having said that, let me now turn to defending the move *from* the fact that only a jerk would say "I recommend that you do this" in these cases *to* the claim that it is

[1] I take this claim about James to follow clearly from the described facts in his case. Laura's case is more subtle; perhaps further details matter, such as how tired she is and how much she has had to deal with sexist comments like these. All I need, to use the case to argue against the Common View, is

false that, morally, each agent ought to take their supererogatory option. My second response to the objector is as follows. I agree with the objector that sometimes only a jerk would recommend to someone that she do the thing that she is morally required to do. And I agree with the objector that sometimes only a jerk would recommend to someone that she do the thing that, morally, she ought to do. But in what kinds of cases is this true? Suppose that someone is morally obligated to do something difficult; she is trying to do it, but she's not really succeeding; she is clearly conscious of her obligation; and no good would come from urging her to do what she is already trying to do, except to make her feel terrible. For example, if one saw one's friend struggling to be patient with her difficult five-year-old son, no good would come of recommending that she be patient; only a jerk would do this. Or suppose that someone faces a difficult choice over whether to do the right thing; he has struggled over his choice, and decided with some regret to violate his obligation; suppose a friend knows that she cannot persuade him to change his mind and that recommending the right action would do no good. For example, if one's friend has decided not to confess to a minor fraud at work, the confession of which would cause the financial ruin of his family, leaving them destitute, then although he has chosen wrongly, it may be that only a jerk would say "turn yourself in." These are indeed two kinds of cases in which only a jerk would recommend to someone that she do the thing that, morally, she ought to do. These are cases in which the agent already knows that, morally, she ought to do it, and she is either struggling to do it, or she has already decided for sure against doing it. These cases are very different from the cases of James and Laura. Consider James in particular. He is not struggling with a difficult choice. And, if he starts saying "no" to playing with Kenny, James is not making any kind of mistake. My response to the objector is that while it is true that *sometimes*, morally, a person ought to do something, although only a jerk would recommend that action, this phenomenon arises in a particular kind of case, and the cases of James and Laura are not this kind of case. When we see that only a jerk would recommend that James or Laura take their morally best option, we have no reason to think that *nevertheless* that option is recommended by morality.[2]

---

that there is a version of Laura's case in which it is false that, morally, she ought to kindly explain to Mark why what he said was awful.

[2] A different objection goes as follows: "It is not obnoxious (or jerky) to say 'morally, you ought to do this' because saying that does not recommend the action; it simply points out something relevant to whether to do it; saying 'morally, you ought to do this' is not the same as saying 'all things considered, you ought to do this.'" This objection misunderstands the argument I make in section 4. My argument is: Only a jerk would recommend taking this option; the recommendations of morality are not recommendations that only a jerk would make; therefore, morality does not recommend taking this option. My argument thus does not rely on the claim that it would be obnoxious (or jerky) to say "morally, you ought to do this" (though I do in fact think that this would be obnoxious, and false, in these cases); rather, my argument relies on the claim that it would be obnoxious to say "I recommend that you do this" or simply to offer the exhortation, "do this."

8 Making sense of morally, you ought to do it

In this section, I will outline the first part of my proposed alternative view, on which the Common View is false. (The first part of My Proposed View concerns morality; the second part of My Proposed View concerns prudence.) The view I develop acknowledges that it is sometimes true that, morally, a person ought to do a particular thing and acknowledges that some *ought* facts are *moral facts*, as in the cases of Aaron, Donna, Georgia, Isaac, and Nicole. Here is my alternative proposal:

> My Proposed View – Part 1: An agent is such that, morally, she ought to φ *just in case*
>
> all things considered (in light of all of her reasons), she ought to φ *and* there are some moral considerations in favor of φing that centrally explain its being the case that she ought to φ.

It follows from My Proposed View that every morally required option is one that, morally, the agent ought to take. This follows because if an option is morally required, then the agent ought to take it, all things considered. (In saying this, I am assuming that moral requirement is overriding.) And if an option is morally required, then there are some moral considerations in favor of taking it that centrally explain its being the case that the agent ought to take it, all things considered. So, on My Proposed View, it is true that, morally, Aaron ought to go to the poetry reading, and it is true that, morally, Donna ought to keep Ellen under consideration.[3]

The cases involving supererogatory options that I've discussed in this chapter show that the following claim is true:

The Supererogation-Ought Claim:

This objection brings out something important about the sense in which morality makes recommendations. The recommendation is simply the recommendation *of an action*; the recommendation is not a claim about the moral properties of an action. For example, if it is true that, morally, an agent ought to dance, then morality recommends *dancing*.

[3] Note that it's clearly true that *if* an agent is morally required to do something, *then* she is such that, morally, she ought to do it. So it's important that My Proposed View accommodate this truth.

Consideration of this truth can explain why My Proposed View includes the condition that there are *some moral reasons* that *centrally explain* its being the case that one ought to do something, all things considered. Compare My Proposed View to the following view: An agent is such that, morally, she ought to PHI just in case, all things considered, she ought to PHI, and *all the reasons that explain* why she ought to PHI are *moral reasons*. Some morally required actions would not meet this condition, because sometimes when an action is morally required, it also has other decisive reasons in favor of it: an action might be both a morally required promise-keeping but also necessary to save the agent from wasting a great deal of money. In that case, some of the reasons in favor of the action, which explain why it ought to be performed, are not moral reasons. So the comparison view would not imply that, morally, the agent ought to perform this morally required action. By contrast, My Proposed View does accommodate the fact that, morally, the agent ought to perform this morally required action.

For related discussion, see section VII of my "Morally Permissible Moral Mistakes" (*Ethics*, 2016).

> Some (but not all) supererogatory options are such that, all things considered, the agent ought to take them. (And some of these are such that the moral considerations that make them morally good to take also centrally explain its being the case that they ought to be taken, all things considered.)

Among moral philosophers who acknowledge the existence of the supererogatory, there has not been adequate appreciation of the fact that sometimes an agent has one or more supererogatory options and the agent ought to take one of them, while sometimes an agent has one or more supererogatory options and yet it is not true that the agent ought to take one of them. The cases of Georgia, Isaac, and Nicole exhibit the first phenomenon: Georgia ought to visit her neighbor for ten minutes, Isaac ought to volunteer to walk the Muslim students home from prayer, and Nicole ought to study the YouTube videos; these are supererogatory options that the agents ought to take. (My considered view about these cases is not that the described details of the cases settle that these *ought* claims are true. Rather, I take the more modest view that the details of these cases are *consistent with* the truth of these *ought* claims. My view is simply that there are versions of the cases of Georgia, Isaac, and Nicole in which these *ought* claims are true.) The cases of James and Laura exhibit the second phenomenon: while both James and Laura have supererogatory options available to them, neither of them ought to take a supererogatory option. (My considered view is simply that there is a version of Laura's case in which it is not the case that she ought to take her supererogatory option.)

The Supererogation-Ought Claim is true because moral considerations continue to have force within the realm of the morally permissible; they can settle that, all things considered, one ought to do something, although it is not morally required. This fact has not been appreciated by philosophers.[4] Taking My Proposed View and the Supererogation-Ought Claim together, the following claim follows:

The Supererogation-Morally-Ought Claim

> Some supererogatory actions are such that, morally, the agent ought to perform them. Some of these are the agent's morally best option; some are not.

[4] I argue for the Supererogation-Ought Claim (though not by that name) in my "Morally Permissible Moral Mistakes" (*Ethics,* 2016) and my "Morality Within the Realm of the Morally Permissible" (*Oxford Studies in Metaethics*, 2015).

In "Morally Permissible Moral Mistakes," footnote 18, I briefly considered a view along the lines of My Proposed View in the current chapter, applied to "what one morally ought to do," and I rejected the view. Though in the current chapter I don't use the locution "what one morally ought to do" (speaking instead of "what, morally, one ought to do"), the view I offer in the current chapter might naturally be applied to that similar locution. Thus, I now think my earlier footnote 18 was mistaken (though the broader point at issue in that footnote was not mistaken – the broader point was that the central lessons of that paper were not already well understood).

The cases of Georgia and Isaac are cases in which, morally, an agent ought to take a supererogatory option which is their morally best option. The case of Nicole is a case in which, morally, an agent ought to take a supererogatory option, but it is not her morally best option.

On My Proposed View, while sometimes it is true that, morally, an agent ought to take her morally best option, in many cases, an agent's morally best option is not such that, morally, she ought to take it. We saw this in the cases of James, Laura, and Nicole. None of these agents is such that, morally, they ought to take their morally best option. Furthermore, in many cases, there is no option, not even a disjunctive option, such that, morally, the agent ought to take that option. We saw this in the cases of James and Laura. The options that James and Laura ought to take, all things considered, are not centrally supported by moral reasons.

9 Prudence does not always recommend the option that is best for the agent

In this section, I argue against the part of the Common View that is about prudence. My argument is parallel to my argument about morality, and the alternative view I develop is parallel.

The Common View makes the following claims:

> (***) Prudentially, an agent ought to do something, just in case doing it would be best for the agent.
>
> (****) Prudence recommends that an agent do something just in case doing it would be best for the agent.

There are plenty of cases in which prudence does recommend taking the option that would be best for oneself. Suppose a pipe has burst in your basement, and it is spewing water everywhere and starting to flood the basement. The best option for you is to go down there right away, get soaked in water, and put a stop to the flooding as quickly as possible; it will be worse for you if you delay. And indeed, prudence recommends that you go down right away. Suppose you have a regular workout schedule in which you go to the gym every Monday morning. You wake up on Monday morning, and are deciding whether to go to the gym. The option that would be best for you is to go to the gym, and indeed prudence recommends that you go to the gym. In each of these cases, the option that is best for the agent is indeed such that, prudentially, she ought to take it.

However, there are cases in which it is not true that prudence recommends performing the action that is best for the agent. Consider Aaron, who is deciding between going to the poetry reading he promised to attend and seeing a Lakers game he would greatly enjoy. The option that is best for Aaron is going to the game. However, prudence does not recommend that Aaron go to the game. Prudence doesn't recommend that Aaron fail to act as morality requires. It's just not true that, prudentially, Aaron ought to go to the game. Consider Isaac, who is deciding whether to volunteer to help keep Muslim students safe, which would involve going out in the cold rain. The option that is best for Isaac is staying warm and dry inside, but prudence does not recommend that Isaac stay inside, and it is not true that, prudentially, Isaac ought to stay inside.

Prudence doesn't recommend that Isaac refuse to do something to help others, even at some cost to himself.

Consider also the following case, which is a fictionalized version of real events (involving a non-philosopher perpetrator):

> Three students have accused a famous professor, Professor X, of sexual harassment. Professor X has responded by suing the students for defamation. The case is widely reported in the news. Another professor says the following about the case: "It is exceedingly unlikely that these three students are lying about what happened to them. So Professor X's lawsuit against them seems to be an effort to intimidate them, though a risky one, as the lawsuit may well dig up other aspects of his past. Given that Professor X's academic career is over unless he prevails, taking this gamble is something that, prudentially, Professor X ought to do, even if he's guilty."

What should we make of these remarks? The speaker claims that prudentially, Professor X ought to sue the students and thus that prudence recommends that Professor X sue the students. But surely both of these claims are false. Notice that the speaker takes it to be exceedingly likely that Professor X has sexually harassed these three students and is now subjecting them to a difficult and stressful defamation lawsuit *on top of* having harassed them. This is the kind of thing that has a high likelihood of driving these students out of their chosen field of study, and some likelihood of driving them into depression or suicide. Surely it is not true that, prudentially, one ought to continue to abuse the people one has sexually harassed by suing them for getting up the courage to tell the truth about what one did. Surely prudence does not recommend such a thing. Prudence does recommend self-preserving activities like eating right and going to the gym. Prudence does not recommend deeply immoral behavior like suing the women that one has sexually harassed to try to scare them into withdrawing their complaints.

The cases of Aaron, Isaac, and Professor X are cases in which prudence does not make any recommendation at all, although the agent has an option that is better for the agent than an alternative. There are also cases in which prudence *does* recommend performing an action that is good for the agent, and yet prudence does not recommend performing the action that would be best for the agent. Consider the following:

> Olive will fail out of school unless she passes tomorrow's test, but she will have to miss the best party of the year if she spends tonight studying. She knows how she could cheat on the test in a way that would not be discovered. Olive could (i) go to the party and cheat on tomorrow's test, (ii) spend tonight studying and miss the party, or (iii) go to the party and then fail the test.

In this case, the option that is best for Olive is option (i), go to the party and cheat on the test. This would enable her to both pass the test and enjoy the party. However, prudence does not recommend that Olive take option (i). Rather, prudence recommends that Olive study tonight. Prudentially, Olive ought to study tonight. This is a case in which prudence makes a recommendation, but it does not recommend the option that is best for the agent.

The cases of Aaron, Isaac, Professor X, and Olive show that the Common View is false. Or, more modestly, my discussion of these cases shows that an alternative view of prudence can be held: if you do not agree with my claims about these cases, I hope they start to show what an alternative view to the Common View would look like and why it has some plausibility. In the next section, I spell out the alternative view.

10 Making sense of prudentially, you ought to do it

We can deny the Common View while acknowledging that it is sometimes true that, prudentially, an agent ought to take a particular option.

> My Proposed View – Part 2:
>
> An agent is such that, prudentially, she ought to φ *just in case* all things considered (in light of all of her reasons), she ought to φ *and* there are some prudential considerations in favor of φing that centrally explain its being the case that she ought to φ.

According to My Proposed View, sometimes an option that is best for the agent is such that, prudentially, the agent ought to take that option. (We see this in the case of going to the gym.) But often the option that is best for the agent is not recommended by prudence, and it is not true in such cases that, prudentially, the agent ought to take that option. (We see this in the cases of Aaron, Isaac, Professor X, and Olive.) In many cases, there is no option, not even a disjunctive option, such that, prudentially, the agent ought to take that option. (We see this in the cases of Aaron, Isaac, and Professor X.)

11 Objections from other *oughts*

In this section, I will consider two objections from consideration of the existence of other purported *oughts*. The first objector addresses me as follows:

> Consider the following truths:
>
> Grammatically, one ought to refrain from ending a sentence with a preposition.
>
> Football-wise, the cornerback ought to tackle the receiver when he is running with the ball.
>
> Etiquette-wise, one ought to use a smaller fork for salad and a larger fork for one's main course.
>
> Legally, one ought to carry one's driver's license whenever one drives a car.
>
> It looks as though there are quite a few *oughts*: an *ought* of grammar, one of football, one of etiquette, one of the law, and so on. Here is a dilemma: either you hold that there are not any distinctive *oughts* like this and apply your account generally to all apparent distinct *oughts*, or you hold that the purported moral *ought* and the purported prudential *ought* are special. If you take the first horn of the dilemma, you face counterexamples: football, etiquette, and the law sometimes recommend actions that one should not take, all things considered. (For example, given the risk of concussions, both to oneself and the other player, one should probably never tackle anyone.) If you take the second horn of the dilemma, you face the burden of explaining why the apparent moral and prudential *oughts* differ from these other *oughts*.

I choose the second horn of this dilemma. Morality and prudence are special. Unlike football, grammar, etiquette, and the law, morality and prudence have a particular feature:

> Morality does not give bad advice.
>
> Prudence does not give bad advice.

This distinguishes morality and prudence from the other sources of advice.[1]

Would I generalize my arguments, and My Proposed View, to these other purported *oughts*? I would not, simply because it is not at all plausible that these other sources of advice never give bad advice.

Consider now another objection:

> It may sound strange to say that *prudentially*, Aaron ought to ditch Bill's poetry reading and go to the Laker game. But it sounds fine to say that *self-interestedly* or *selfishly*, that's what Aaron ought to do. But since prudence is simply the same as self-interest, this shows that it's actually true that, prudentially, Aaron ought to ditch Bill's poetry reading and go to the Laker game.

In response to this objection, I am happy to grant that, as far as selfishness goes, Aaron ought to ditch Bill's poetry reading and go to the Laker game, or, we might say, taking only his own interests into account, that is what Aaron ought to do. What this brings out is that "prudence" is not a name for selfishness or for self-interest. Nevertheless, on my view, prudence is indeed distinctly concerned with the agent's own well-being. But prudence is concerned with the responsible and warranted pursuit and protection of one's own well-being.[2] That is, prudence is concerned with the pursuit of one's own well-being that is supported by one's reasons.[3]

---

[1] Susan Wolf, "Moral Saints," *Journal of Philosophy* 79 (1982): 419–439, discusses a "moral saint," someone who always does the morally best thing she could do. Wolf discusses *whether* one should want to be a moral saint; her view is that one should not. I am arguing that not even *morality* urges such a life upon us. Morality urges morally good action only when, all things considered, it is the thing to do: morality sometimes urges us to do the morally best thing we could do, but it often does not.

[2] There is some use of "prudent" to mean simply wise or responsible; that is not the use of "prudent" or "prudence" in this chapter.

[3] An objection from the opposite direction goes as follows: the Common View would not hold that, prudentially, Professor X ought to sue his students, because it makes one's life go worse when one does something deeply immoral, and it's immoral for him to sue his students. In response to this objection, let's recognize a *narrow* and a *broad* sense of what's good for a person. Suppose it's true that being immoral makes one's life overall worse; that is true in the broad sense of what's good for a person. But we still should acknowledge that sometimes a person does something deeply immoral because it is better for him; this is true in the narrow sense of what's good for a person. The plausible version of the Common View – on which morality and prudence sometimes give conflicting recommendations, and prudence always recommends what's best for the agent – uses this narrow sense of what's good for a person. So the objector is incorrect.

12 Summary

I've argued that the Common View is false, and I've proposed an alternative view. On My Proposed View, the considerations that make something the morally best thing to do are distinct from the considerations in virtue of which morality *recommends* one action rather than another. Something is a morally good thing to do because of the other-regarding considerations that tell in favor of doing it. Whether something is a morally good thing to do is insensitive to how much it burdens the agent. (Or perhaps the fact that an action burdens an agent more severely makes it a morally better thing to do!) But what morality *recommends* does take into account burdens on the agent, just as what morality *requires* takes burdens into account. Similarly, on My Proposed View, what prudence *recommends* is sensitive to all the agent's reasons, not just her self-interested reasons.

On My Proposed View, it is sometimes true that, morally, an agent ought to take a particular option, and it is sometimes true that, prudentially, an agent ought to take a particular option. But each of these claims is true only if, all things considered, the agent ought to take that option. On the view I've offered, these claims never conflict: it is never the case that, morally, an agent ought to do something, while prudentially, she ought refrain from doing it. However, it can be true that *both* morally, an agent ought to do something, and also, prudentially, she ought to do the same thing. This will be true whenever moral and prudential considerations both provide central explanations of why an agent ought to do a particular thing.

My Proposed View does not hold that there are three distinct *oughts*, one moral *ought*, one prudential *ought*, and one all-things-considered *ought*. Rather, it is the all-things-considered *ought* that is at play throughout the phenomena we have discussed. Some *ought* facts are *moral* facts in that they are centrally explained by moral consid-

erations.[1] Some *ought* facts are *prudential* facts in that they are centrally explained by prudential considerations.[2]

[1] An objector might react to my argument as follows:

You're argued that there is no distinctively moral *ought*. But there is a *pro tanto* moral *ought*: whenever someone has a *pro tanto* moral obligation to do something, one can truly say "morally, she ought to do this" even though often it is false that, all things considered, she ought to do it.

Let's grant to the objector that one can truly say "morally, she ought to do this" whenever someone has a *pro tanto* moral obligation. However, the existence of a *pro tanto* moral *ought* would not challenge my argument. In this chapter, I have argued against a moral *ought* such that, morally, an agent ought to something, just in case it is the morally best thing she could do. A *pro tanto* moral *ought* is not the kind of *ought* I am arguing against, because it will often be true that, morally, an agent ought (*pro tanto*) to do something although doing it is not her morally best option.

[2] There is a feature of my view that might seem peculiar: from the fact that, morally, S ought to φ, and the fact that ψing is an alternative to φing, it does *not* follow that, morally, S ought to refrain from ψing. For example, morally, Nicole ought to watch the YouTube videos; taking the ASL class is an alternative to watching the YouTube videos, but it is not the case that, morally, Nicole ought to refrain from taking the ASL class: there is no moral problem with her doing so. But this feature of my view will seem peculiar only if we take there to be a moral *ought* at play in these facts. That is not right. There is simply the all-things-considered *ought* at play in these facts. And it does follow from the fact that, morally, Nicole ought to watch the YouTube videos (and that taking the ASL class is an alternative) that Nicole *ought to refrain* from taking the ASL class: but it is simply true that, all things considered, she ought to refrain from taking the ASL class.

This apparent peculiarity helps to illustrate that on my view, it is really not the case that there is a moral *ought*.

13 Why this matters

Why is it important to see that My Proposed View is true and that both the Common View and the Naïve View are false? Moral philosophers' failure to see the truth of My Proposed View partly explains, and is partly explained by, their being caught in the grip of a false picture of moral requirement and of the nature of the realm of the morally permissible.

This chapter is part of a series of papers[1] in which I have been arguing that philosophers mistakenly endorse the following characterization of moral requirement:

The False Characterization:

> An agent is *morally required* to do something just in case, all things considered, she ought to do it, and the reasons that explain why it ought to be done are *moral reasons*.

This characterization is incorrect, however, because of the truth of the following claim (which I introduced in section 8):

The Supererogation-Ought Claim:

> Some (but not all) supererogatory options are such that, all things considered, the agent ought to take them. (And some of these are such that the moral considerations that make them morally good to take also centrally explain its being the case that they ought to be taken, all things considered.)

Consider the case of Georgia. In this chapter, I focused on the claim that, morally, Georgia ought to visit her sick neighbor Harriet. But it is also true that, all things considered, Georgia ought to visit Harriet; and moral reasons explain why. Georgia is not morally required to visit Harriet; thus, this case shows the False Characterization to be false.[2] If in this case, Georgia fails to visit Harriet, then Georgia makes a *mistake* (she acts as she ought not to act, all things considered); it is a *moral* mistake (in that the reasons she ought not to act this way are moral reasons), and yet it is *morally*

[1] See footnotes 13 and 22.

[2] Note that I only need the claim that there is some version of Georgia's case in which, all things considered, she ought to visit Harriet. I do not need the claim that the details of the case make it obvious that, all things considered, Georgia ought to visit Harriet. More generally, what I need is the claim that *sometimes* a supererogatory action is the thing to do, the thing that is the correct choice in

permissible. In my recent series of papers, I have been arguing for the importance of recognizing that some actions are *morally permissible moral mistakes*.

Four things are at stake in properly understanding these moral issues. First, it is important that we not endorse the False Characterization of moral requirement; we should not misunderstand the name of moral requirement. Second, it is important that we correctly understand the way that reasons function within the realm of the morally permissible: among morally permissible options, moral reasons continue to have sway, and they can settle that, all things considered, one ought to take a supererogatory option. We must correctly understand the normative status of supererogatory options: some are such that, all things considered, they ought to be taken, but some are not. As I have argued in this chapter, some supererogatory options are such that *morally*, they ought to be taken, but some are not. Third, if we are caught in the grip of the false picture, then we will misunderstand the force of *moral arguments* against behaving in certain ways: we will assume that every moral argument against a way of behaving shows that behavior to be *morally wrong*. But it turns out that a moral argument may merely show a way of behaving to be a moral mistake without showing it to be morally wrong. Fourth and finally, once we abandon the false picture and recognize the existence of *morally permissible moral mistakes*, we can recognize the possibility of new moral views that we have not considered. For example, consider the view that it is a *moral mistake* to eat meat and that eating meat is a morally bad thing to do,[3] and yet that it is not morally wrong to eat meat. This view might explain what is otherwise puzzling: that some vegetarians refrain from eating meat for moral reasons, and yet they accommodate meat eating in others. It is not in general morally permissible to accommodate others' wrongdoing; but it may be permissible to accommodate others in committing mere moral mistakes. Another example is given by the view that gamete donation (sperm and egg donation) is a moral mistake, though it is also a wonderful thing to do; in other work, I argue that this view is worth taking seriously.[4]

We might say that the crucial insight is that the category of *moral mistakes* is bigger than the category of *morally wrong behavior*. Moral mistakes are options such that, all things considered, one ought not to take them, and moral reasons explain why they ought not to be taken; some of these are not morally wrong. But we could equally say that what's crucial is to recognize the *flipside category*: the category of options such that, all things considered, one ought to take them, and moral considerations explain

light of one's reasons: sometimes, all things considered, one ought to perform a supererogatory action.

[3] Note that a way of behaving might be a moral mistake without being a morally bad thing to do, so the claim that eating meat is a morally bad thing to do goes beyond the claim that eating meat is a moral mistake. (For example, if Nicole fails to learn any sign language, neither watching the YouTube videos nor taking the class, then she makes a moral mistake in failing to watch the YouTube videos, but she doesn't do anything morally bad. She merely fails to do something morally good that she ought to have done.)

[4] For extended discussion of these two views, see my "Eating Meat as a Morally Permissible Moral Mistake," in *Philosophy Comes to Dinner* (New York: Routledge, 2015) and "Gamete Donation as a Laudable Moral Mistake," in *Oxford Handbook of Population Ethics*, eds. Gustaf Arrhenius, Krister Bykvist, Tim Campbell, and Elizabeth Finneron-Burns (Oxford: Oxford University Press, Forthcoming).

why one ought to take them. This category does not just include morally required options; it also includes some supererogatory options.[5] In this chapter, I've argued that this flipside category is the category of *options that, morally, ought to be taken.*

So now we come to the topic of this chapter. Those who endorse the False Characterization of moral requirement do not see the *flipside category* as a distinct category from the category of morally required options. Thus, when seeking to understand the category of *options that, morally, ought to be taken*, they see only two plausible views: the Naïve View, which equates *options that, morally, ought to be taken* with morally required options, and the Common View, which equates *options that, morally, ought to be taken* with morally best options. It is only in rejecting the False Characterization that one can see a distinct third category: as I've argued in offering My Proposed View, the *options that, morally, ought to be taken* are those options that meet the following condition: all things considered, the agent ought to take the option, and moral considerations explain why it ought to be taken. This includes all morally required options, but it also includes some supererogatory options.[6]

[5] If there are any *suberogatory* actions, then this flipside category also includes some (or all) refrainings from suberogatory actions. Supererogatory behavior is what is morally good to do but not morally required; suberogatory behavior is what is morally bad to do but not morally wrong. It is disputed whether any behavior is really suberogatory, even among those who believe in the supererogatory. For further discussion of how the suberogatory fits into this flipside category, see my "Morally Permissible Moral Mistakes," section V.E. For arguments for and against the suberogatory, see Julia Driver, "The Suberogatory," *Australasian Journal of Philosophy* 70 (1992): 286–295 and Hallie Liberto, "Denying the Suberogatory," *Philosophia* 40 (2012): 395–402.

[6] For helpful comments on drafts of this chapter, I thank Tyler Doggett, Stephen Finlay, Peter Graham, Alex Guerrero, John Horty, Elinor Mason, Douglas Portmore, Mark Schroeder, Ted Sider, Kurt Sylvan, my spring 2017 graduate seminar at Princeton University, and audiences at Brown University; Florida State University; Princeton University; Purdue University; University of Edinburgh; University of Reading; University of Texas, Austin; and University of Southern California.

30. Practical reason and the second-person standpoint

Stephen Darwall

1 Introduction

Almost fifty years ago, Thomas Nagel's *The Possibility of Altruism* focused philosophers' attention on the perspective or standpoint from which reasons for acting are registered in deliberation (1970). Nagel distinguished between what he called the "personal standpoint" and the "impersonal standpoint," corresponding roughly to what we today might more usually call "first-person" and "third-person" perspectives, respectively. Reasons for acting that can only be appreciated first-personally, from the "personal standpoint," Nagel called "subjective reasons." And those that can be weighed as reasons to act third-personally, from an "impersonal standpoint," he called "objective reasons."[1] Such considerations as that an action would benefit *me* (or *us*, for that matter) or bring about something *I* (or *we*) desire are thus subjective reasons. By contrast, the facts that an action would benefit *someone*, or even that it would benefit S.D. (the person I happen to be) or bring about something S.D. desires, are objective reasons, since they can be appreciated third personally.[2]

Nagel's object in drawing this distinction was to make an argument that all normative reasons are ultimately objective, comprehensible from an impersonal, third-person standpoint. At the time Nagel wrote, however, philosophical orthodoxy held the opposite. Under the influence of Davidson on desire-belief action explanation and social scientific models of "economic rationality," it was widely maintained that any normative reason for acting must be concerned with furthering *the agent's* interest or satisfying *her* desires or preferences (Davidson 1963).

Nagel's argument was extraordinarily ambitious. He sought to argue not just that considerations of impersonal benefit are *among* reasons for acting, with subjective reasons of personal benefit providing reasons also. His object was to show that no genuine reason for acting could be non-derivatively subjective. To believe otherwise, he argued, is to be committed to a kind of "practical solipsism" whose theoretical analogue is well outside the philosophical mainstream.[3] To fully appreciate that one is

[1] Nagel construed normative practical reasons as predicates ("R") applying to actions such that when they hold any agent ("p") has pro tanto reason to perform them. Nagel assumed, tendentiously, that all reasons for acting are reasons to promote some state of affairs, so that what he wrote was actually that any normative reasons can be expressed by "a predicate R, such that for all persons p and events A, if R is true of A, then p has prima facie reason to promote A" (Nagel 1970: 90). As Nagel put it, a reason is a *subjective reason* if its "defining predicate R contains a free occurrence of the variable p" (Nagel 1970: 90).

[2] Although, of course, the *de dicto* fact that it would benefit the person I happen to be is a subjective reason.

[3] Although Nagel characterized his argument as Kantian, it was more firmly rooted in G. E. Moore's argument against egoism (Moore 1993: 148–151).

simply one "among others equally real" is to be committed to taking one's own desires and interests as reasons only because they are *someone*'s (Nagel 1970: 14).

Few philosophers were convinced to accept this ambitious conclusion. For one thing, it is hard to see how a rational egoist, who holds that each and every agent should take her own interests as reasons to act, can plausibly be accused of failing to take seriously the reality of other persons. Her view is not about herself alone but about any rational agent. She holds that each agent is the sole object of her own rational concern.

Nevertheless, the assault that Nagel began on the widespread notion that normative reasons must always be grounded first-personally in the agent's own desires and welfare ultimately bore fruit (Darwall 1983; Bond 1983; Pettit and Smith 1990; Parfit 1997; Scanlon 1998; Dancy 2000; Parfit 1997, 2011). Assume, for example, that Davidson is right that all actions can be explained teleologically by the agent's desires and beliefs. Whenever someone acts intentionally, there must be something she desires that she believes she can accomplish by so acting. In one sense, then, reasons for her action will be grounded in her desires. But this only means that explanatory, *motivating* reasons for her action somehow involve her desires, not that any *normative* reasons for her so to have acted must be. From Davidsonian considerations about action explanation alone, nothing whatsoever follows about normative reasons agents have *to* act.

Nagel himself showed as much with his distinction between "motivated" and "unmotivated desires" (Nagel 1970: 29–30). Desires can be "motivated" or had for reasons themselves. If someone wants something, we can ask what his reasons are for wanting it.[4] So far is it from being the case that the teleological explainability of action by desire shows that normative reasons are grounded first-personally in the agent's desires, indeed, that when we reflect on the phenomenology of desire itself, it can begin to seem mysterious how the fact that an agent desires something can provide any normative reason to act at all.

Philosophers these days generally hold that normative reasons for acting (or for desiring something) are *facts* that count in favor of or that cast a favorable light on whatever it is that they are reasons for and that an agent can take account of in thinking about what to do or desire (Darwall 1983; Parfit 1997; Scanlon 1998; Dancy 2000). Usually when one acts as one desires, however, one is not moved by an awareness of the fact *that one desires*. To desire is to have one's attention drawn to the *object* of desire and features of or facts about the object, which one regards as reasons to desire it, rather than on the fact that one desires (Darwall 1983: 37). Accordingly, when one is motivated by a desire to act, there will normally be facts about the desire's object that one will take as reasons so to desire and act, and these need not be dependent on the fact of one's desire in any way.

[4] When philosophers talk about motivating reasons for an action, they sometimes refer, as Davidson did, to the agent's mental states, his desires and beliefs that explain the action teleologically. But "motivating reason" can also refer to what are frequently called *the agent's reason*, namely considerations the agent regarded as (normative) reasons and *for* or *on which* he acted (Darwall 1983: 32). For an attempt to clarify these different usages, see Darwall 2003, esp. 442–443.

This is the feature of agency that Pettit and Smith called the "backgrounding of desire" (Pettit and Smith 1990; see also Darwall 1983; Bond 1983; Scanlon on "desire in the directed attention sense" Scanlon 1998: 39; Dancy 2000). The situation is exactly analogous when we reason from beliefs. Although any reasoning culminating in belief must begin with prior beliefs, it typically proceeds from *what* one believes, rather than from the (first-personal) fact that one has certain beliefs. Clearly what one believes need be nothing subjective (in Nagel's sense) or first personal at all. Indeed, it is consistent with the fact that theoretical reasoning always assumes prior beliefs that the assumed beliefs are *never* subjective or first personal. They might always be (believed to be) objective facts that one appreciates from a third-person point of view. Even when we base our beliefs about the world on appearances, though these are necessarily from one's point of view, nothing first personal need be part of the *content* of the appearance. If something looks red to me, it looks *to be red*, not *to be red to me*.

Similarly with desire, standardly, when we desire something, some state of affairs involving that thing seems good or desirable, something there is reason *to* desire and, therefore bring about (see, especially Scanlon on desire and buck-passing, 1998: 39, 95–97).[5] Agents motivated by such desires will thus take desirable-making facts and features of that state of affairs as reasons to act. It follows that their reasons will be subjective or first personal just in case those features and facts can only be expressed with a first-person pronoun.

Of course, this is often the case. If I want the experience of a hot shower on a cold day, what I want is that *I* have this experience. But many desires are not first personal in this way. Desires that structure ordinary forms of benevolent concern or care for others obviously are not. Some of our strongest desires, those of parents for their children's welfare, or of friends for one another, are not self-referential in their content. Though being my children is part of the explanation of why I care so much for Julian and Will, my caring is nonetheless for them *de re*. In loving them, I see them as having a value that makes them worthy of anyone's concern (Darwall 2002: 45–49).

There is even, indeed, a form of concern for oneself that is the analogue to benevolent concern for others that is not essentially first personal. Benevolent concern for oneself *de re* rather than *de se*, for the person one is rather than for oneself whoever one is (and only on that condition), is no less third-personal than is benevolence for others (Darwall 2002: 49).

In cases like these, the reasons for which agents desire and therefore act are third personal rather than first personal. Of course, this just means that there are third-personal *motivating reasons*, in the sense of *agent's reasons*. Nevertheless, when agents act for such reasons, they, at least, are taking it that these third-personal considerations are valid normative reasons for them so *to* act.

I take it that you, dear reader, are like the agents I have been describing in that you take both first-personal and third-personal considerations like these (and many

[5] But see Schapiro 2009. See also Tenenbaum in this volume.

others) to be normative reasons for action. Even if you haven't thought about it as a matter of theory, I would bet that you believe that there are both first- and third-personal normative reasons for acting, as, I conjecture, would be evidenced by the best interpretation of your actions. You do some things for first-personal reasons, say, because there are certain things you want to experience (yourself) or because, as we sometimes say, "it's just something you want to do." But there are other things you think have a value or importance that is not merely personal but that could be called "impersonal" to signal that these things are valuable from a third-person point of view. Values that give our lives meaning are like that (Darwall 1983; Wolf 2012). Indeed, much of the contribution to our personal welfare that meaningful activities make comes from the deeply satisfying appreciation of their (impersonal) value and importance that fully enjoying them can involve (Darwall 2002: 73–104).

2 Second-personal reasons

That some reasons for acting are first personal and some third personal is thus arguably implicit in what we might plausibly regard as common sense about practical reason. My main object in what follows is to argue that the same is true about what I shall *second-personal reasons*, normative reasons for acting that implicitly involve the perspective you and I are in right now and that I made explicit in the last paragraph when I addressed you by saying "you, dear reader." Second-personal reasons are reasons that are distinctively *addressed* person-to-person. When we act on them, we implicitly recognize a shared second-personal *authority* we assume that persons have to address these reasons to one another.

Suppose, for example, that you and I encounter each other on the street, and I ask you for directions. Obviously, I hope that you will treat my asking you as a reason to give me the directions. If you do so *because I asked*, you will treat my asking as *giving* you a reason to give me directions, not simply as making evident or "triggering" a reason that existed independently of my asking (Enoch 2011). To be sure, answering requests might have an impersonal value that would generate third-personal reasons, but asking/being asked is itself a second-personal relation, and the reason "that I asked you" is a second-personal reason. Were you to treat my asking as such a reason, you would be implicitly acknowledging my authority or standing to ask this, though not perhaps to expect it, of you. Your answering would itself be second personal in the sense that it would reciprocate the question and relate to me as someone who may ask it of you.

There might of course have been other relevant first- and third-personal reasons. Seeing me disoriented, you might have thought you had a third-personal reason to help me independently of being asked. Or perhaps you had some other first-personal reason; maybe you wanted to have the experience of giving directions because you are in rehearsals for a play in which your character has similar lines. Of course, even this first-personal reason would involve your pretending to treat my asking (you) as a second-personal reason.

A natural thought might be to try to reduce this second-personal reason to more fundamental third-personal reasons of benefits to me *de re*, S.D., the person I am. But you don't need to know anything about whether having the directions would be good for me to think you have a reason to give them just because I asked you for them.[1] But what about the third-personal reason that giving them to me would be doing

[1] Of course, this reason might be overridden or defeated.

something that S.D. wants? However, now ask yourself, does the fact that I *asked you* make no difference beyond giving you evidence of my desires? We can easily imagine cases, I think, where it *does* make a difference. Think, for example, of cases of romantic or even friendly approach. Or think, indeed, of the current situation, where I have just asked you to think of a case. Even if it were common knowledge that people never ask for things they do not want, that would not mean that their asking does not provide an additional reason.

Since "you" is a second-person pronoun, "that I asked you" is a second-personal reason in a straightforward, Nagelian sense that it is only expressible second personally. However, that is not necessary for something to be a second-personal reason in a broader sense that I will be concerned with beginning in the next section (and that features in my work more generally, for example, in Darwall 2006, 2013a, 2013b). There I will be calling attention to a class of reasons that presuppose relations of mutual accountability and authorities to make claims and demands of one another and where acting on these reasons is always a move *within* such reciprocal relations. Any such reason is a second-personal reason in the extended sense that the mutual accountability they presuppose consists in forms of justified second-personal address.

3 Deontic second-personal reasons

Presently, I shall be arguing that the deontic moral ideas of obligation, duty, right, wrong, wronging, and rights are all second personal and that reasons employing these concepts are second-personal reasons in the extended sense I just mentioned. These are all considerations that we register as reasons in implicitly relating *to* one another, and ourselves, as mutually accountable moral agents. As a way of transitioning to that argument from the present example, consider the related idea of *consent*. (Notice, by the way, that I am asking you to do this!)

Consent is second personal in its nature. Suppose (oops, there I go again) that I would like to borrow your copy of Plato's *Republic*. You can consent to that and "let me" have it. But to do that, you have to *give* me your consent, which you can only do by addressing me second personally. In giving me your consent, you relate to me as someone who has the authority to ask for it, and in receiving it, I relate to you as someone who has the authority to give it. Likewise, if having received your consent, I then make use of your copy of the *Republic* (only) because *I have your consent*, this also is a way of relating to you that acknowledges to you that the book is yours and that I may use it only because you have consented to that.

Notice, again, how the second-personal fact of having received your consent affects my normative reasons additionally to any reasons supplied by facts about our respective welfare and wishes.[1] Suppose that you and I don't know each other, but that I know from others' testimony that you likely would not mind were I to take your book for a couple of hours. Having your actual consent obviously affects my reasons for taking it in a way that this knowledge alone would not. To see this, consider recent discussions of sexual misconduct on college campuses. The doctrine that "affirmative consent" is required, or even the less demanding standard that "'no' means no," shows that facts about another's well-being or even about their wishes cannot substitute for the actual second-personal transaction of the giving and receiving of consent.

Consent is an example of what philosophers call a "normative power" (Raz 1972). The exercise of normative powers changes rights and duties directly and not by way of changing some further right- or wrong-making feature. Another example is promising. If I promise to return your copy of the *Republic* by Monday, I undertake an obligation to do so. Moreover, this obligation is a "directed" or "bipolar" obligation *to you*, as is shown by the fact that my promise gives you a claim right to a Monday return

[1] Consent doesn't of course create positive reasons to do actions that are consented to. Rather it disables reasons against that would otherwise consist in the wrongness of the action without consent.

(Darwall 2012). If I do not return your book on Monday then, other things equal, at least, I will not simply have done something wrong. I will have wronged you.

I might have an obligation (pure and simple) to return the book even if neither you, nor indeed anyone, had the right to my doing so, hence even if were not obligated to do so *to anyone*. Similarly, when I gained your consent to borrow your *Republic* in the first place, you gave me a waiver of your right to exclusive use of the book during that period. Before that, you had a claim right to exclusive use, and I was obligated, not just pure and simple, but *to you* not to take it.

Normative powers like promise and consent are powers to directly change rights and the bipolar obligations that are their correlates. They can only be exercised second personally in a reciprocally recognizing transaction in which two parties (the giver and receiver of consent, for example) mutually acknowledge their respective authorities to give and receive consent and thereby change their mutual obligations and rights, and therefore, their reasons for acting in the relevant ways. Similarly, a promise must be made *to* a promisee in a second-personal transaction in which both implicitly acknowledge to one another the promiser's authority to promise and the promisee's authority to accept or reject the promise (see, e.g., Darwall 2011).

Reasons for acting that are generated by the exercise of normative powers are thus second personal. They are created by second-personal relatings, like consent, promise, and request (as in my asking for directions, or in my asking you, dear reader, to consider and suppose various things in this very essay). Acting on these reasons, moreover, is itself implicitly second personal; it implicitly acknowledges second-personal authorities to address (and be addressed) (second-personal) normative reasons for acting and relates *to* others on these terms.

In *The Second-Person Standpoint* (*SPS*) and subsequent work, I argue that the central deontic moral ideas of obligation, duty, requirement, demand, permission, right, wrong, wronging, and rights are all implicitly second personal (Darwall 2006, 2013a, 2013b). These deontic ideas are interdefinable via the following conceptual truths. Something is wrong (either other things equal or all things considered) if, and only if, it violates a moral duty (obligation, demand, requirement) (again, other things being equal or all things considered). Something is permissible if, and only if, it is not wrong. Something wrongs someone if, and only if, it violates a bipolar obligation owed to them. Something is owed to someone by someone else if, and only if, the former has a claim right to it against the latter. Finally, something wrongs someone (violates an obligation to them) only if, but not if, it is wrong (pure and simple). The reason it is not a conceptual truth that if someone does wrong, then there is someone wronged is that it is conceptually possible, at least, for there to be moral obligations (pure and simple) that are not owed to anyone (having a claim right to it). One might think, for example, that it is wrong to destroy things of great natural beauty even if this does not wrong some person. Whether or not this is true in fact, it is surely a conceptual possibility.

This does not mean, by the way, that the fact that an action would be wrong is not itself a second-personal reason, though the fact that it would wrong someone is. Presently, I shall be arguing that wrongness and moral obligation, pure and simple, provide second-personal reasons no less than do facts concerning wronging and the violation of bipolar obligations. This is because these "unipolar" ideas conceptually implicate *culpability*, and that entails (second-personal) accountability to the moral community or to everyone as representative persons, as opposed to accountability to specific individuals as is involved in bipolar obligations and claim rights.

I have argued already that normative powers are second personal in their nature. So also are the bipolar obligations and claim rights they conceptually implicate. Consider what it is to have a claim right against someone or, correlatively, for the latter person to be obligated *to* one as a right holder. If I promise to return your copy of the *Republic* by Monday, I become obligated to you to do so and give you a claim right against me to comply with this obligation. But what do the bipolar obligation and claim right, respectively, consist in?

I argue that they consist in distinctive second-personal normative relations. If you have a claim right against me, then you have an *individual* authority to hold me personally accountable as the very individual to whom I am obligated. This gives you a standing with respect to me and my returning your book that others do not have. For example, you, but not others, can release me from my obligation to you. You can insist on my complying with it in a way others cannot. If I do not keep my promise to you, then you also have distinctive authorities to resent the injury or wronging, to decide whether to seek compensation, to forgive the wronging, and so on. No one can forgive me on your behalf; that is up to you alone.

Bipolar obligations and claim rights thus entail the *individual authority* of the right holder to hold the person obligated to her, the obligor, *to* the obligatory action. The obligee has a distinctive standing to hold the obligor *personally accountable* to her that others do not have. This is part, at least, of what the difference being my being obligated *to you* to return your book by Monday and my being morally obligated pure and simple to return it consists in.

But what about the latter? In what does moral duty or obligation *simpliciter* consist? A central claim of *SPS* is that deontic moral concepts are all second personal in the sense of being conceptually tied to accountability, which, I argue, following Strawson and others influenced by him, presupposes the second-personal authority to address demands (Strawson 1968; Watson 1987; Wallace 1994; Darwall 2006, 2014a, 2014a). The difference between bipolar moral obligations and moral obligation pure and simple is that whereas the former conceptually implicates the obligee's *individual authority* to hold the obligor personally accountable, the latter involves any and every moral agent's *representative authority* to hold the obligor accountable as a representative person (or member of the moral community) (Darwall 2012). Every person has this authority, including the obligor himself, as she must presuppose when, for example, she holds herself answerable through the emotional attitude of guilt. For example,

it is wrong, pure and simple, to destroy natural beauty, even if that does not wrong anyone, if, and only if, destroying nature is something we are accountable for doing, not to anyone in particular but to the moral community or anyone in general.

Strawson famously identified a set of mental states, which he called *reactive attitudes*, that are held from a perspective of implied relationship with or to their objects. Unlike "objective" attitudes, which we hold toward others from a third-person perspective, reactive attitudes, because they implicitly address their objects second personally, necessarily presuppose that their objects are moral agents with the requisite capacities to be held, and, indeed, to hold themselves, accountable. These are what Gary Watson calls "constraints of moral address" (Watson 1987). Examples of the reactive attitudes that are implicated in deontic moral concepts are resentment, moral indignation or blame, and guilt. To have the attitude of blame toward someone, one must view him as a moral agent who is competent to be held thus accountable and, I have argued, to hold himself accountable in his own conscience.

The conceptual link between deontic moral concepts and second-personal accountability goes via the concept of *blameworthiness* or *culpability*. It is a conceptual truth, I claim, that if an act is morally obligatory (wrong, or morally impermissible, not to do), then it is an act of a kind that would be blameworthy to perform were one to do it without excuse. Attempts to understand moral obligation (the deontic sense of the moral 'ought') in terms of the balance of moral reasons fail to capture its distinctive second-personal character through its conceptual connection to culpability.

It is true that philosophers sometimes speak of obligation and blame in other areas than the moral, for example, of epistemic or even prudential "obligations," where, like the moral case, violation may be held to require excuse. However, I am skeptical that what these philosophers have in mind by blame is a genuinely Strawsonian attitude of a sort that can be warranted only if the attitude of guilt is warranted also for its target. Of course, we can "moralize" either epistemology or prudence and think that the Strawsonian attitude of blame and its reciprocal guilt *are* warranted for violating epistemic or prudential "requirements." But that would amount to holding that we have a moral obligation not to violate these epistemic or prudential standards.

Moral obligation's conceptual tie to culpability can be shown with an "open question" argument. We can easily imagine two people who agree that a course of action is what moral reasons most recommend – what it would be morally best to do in that sense – but disagree about whether the action is morally obligatory. One might hold that it is and the other deny this because, for example, the sacrifice to the agent is sufficiently large to make the action *supererogatory*. Suppose, to make it more vivid, that one is an act consequentialist and that the other holds that what consequentialism dictates for this case is "too demanding" to plausibly be morally required. For present purposes, all that matters is that this is a coherent disagreement, not whether the correct *normative* theory of moral obligation would countenance a category of the supererogatory.

It seems clear that such a disagreement is possible. The explanation, I submit, is moral obligation's conceptual tie to accountability, its second-personal character.

What the parties would be disagreed about is whether, in such a case, someone could justifiably be *blamed* for failing to perform the morally recommended action, that is, whether such a failure would be *culpable*.

If this is so, it follows that the deontic moral ideas of obligation, duty, permissibility, right, wrong, and the like are all second-personal concepts. And if such facts are normative reasons for action, they must be second-personal reasons. In acting on such reasons, we implicitly hold ourselves accountable to all persons, ourselves included, for compliance and relate to others and ourselves as mutually accountable moral agents or persons.

4 Morality's authority

Why suppose, however, that such moral facts actually do provide normative reasons for acting? Philippa Foot famously argued that although morality certainly purports to provide such reasons, valid normative reasons for acting are all grounded in the agent's own desires and interests (Foot 1972). If this were so, no deontic, irreducibly second-personal, moral reasons would exist.

Foot was assuming that all normative reasons are first personal, which, as we noted, was a widely held view in the early 1970s. But first-personal considerations of the agent's own desires and interest are hardly self authenticating. The general problem of explaining the "source of normativity" for any normative reasons at all besets first-personal considerations no less than second-personal or third-personal ones. If, consequently, there are reasons to be skeptical of second-personal reasons in particular, they cannot derive simply from the general problem of normativity.

The issue is whether there is any reason to be skeptical of second-personal deontic moral reasons that does not extend to other putative normative reasons also. In a second-personal analysis, moral obligation facts necessarily co-vary with facts about what people can justifiably be held to, facts about what it would be blameworthy to do without excuse. It would seem to follow that there can be deontic moral reasons if, and only if, there can be reasons for the distinctive attitude of blame.

Suppose, however, that we assume that normative reasons *do* exist, not just for action, but also for a wide variety of attitudes, including blame. A further problem might seem then to arise. Suppose we know that it would be wrong for me to refuse to return your copy of the *Republic* and conclude from this that any unexcused failure of return would be blameworthy. It might seem that all this licenses us in concluding is not that returning your book is something there is reason for me *to do* but only something that failing to return it is something anyone (including I) would have reason *to blame*.

But consider the nature of blame as a holding accountable attitude. And imagine blaming me (or me blaming myself) for failing to return your book. Could we coherently be in this state and also accept that I had sufficient normative reason not to return it or that the fact that I was morally required to return it did not give me normative reason to do so? It seems not. It would simply be irrational or incoherent to blame me for doing something one thought I had sufficient reason to do. Blame cannot therefore be warranted, an action cannot be culpable, if there was sufficient reason to do it. Charges of culpability can be defeated either by excuse or by *justification*, that is, there being sufficient normative reason. If I can answer for myself by providing sufficient normative reason, then I do not need an excuse to show my failure to return your book is not

culpable. If consequently, failure to return is wrong, and it would be culpable without excuse, then there cannot be sufficient reason for such a failure.

It follows from moral obligation's conceptual connection to accountability and the nature of blame as a second-personal holding accountable attitude that being morally obligated to do something conceptually guarantees not just normative reason to blame unexcused wrongful action but also sufficient normative reasons not to do what is wrong. If, consequently, there can exist second-personal reasons that justify holding people accountable through reactive attitudes like moral blame, it will follow further that there are second-personal normative reasons for acting and, moreover, that these will have sufficient weight to override reasons of other kinds when they conflict. This does not mean, as it is sometimes supposed, that moral reasons always override other reasons. It means that when these combine to make an action morally *required*, hence blameworthy to omit without excuse, this second-personal deontic moral fact guarantees the presence of conclusive normative reasons for acting as morally required.

5 Conclusion

Appreciating the second-personal character of moral obligation thus enables us to see, not only that second-personal reasons are among the normative reasons we have but that some are especially weighty. Moral obligations give us second-personal reasons that override (or perhaps silence or even pre-empt or exclude [McDowell 1979; Raz 1986]) normative reasons to the contrary. Although this conceptual point is, strictly speaking, independent of any substantive normative claims, I argue in *SPS* that it may nonetheless be viewed as providing a foundation for the normative theory of moral right that Scanlon calls contractualism (Scanlon 1982, 1998; Darwall 2006). Contractualism grounds morality in the idea of justification *to* one another. But this just is the idea of mutual accountability, or what I call "equal second-personal authority" – our equal status as moral persons to make any claims and demands of one another at all. If this is right, then not only will appreciating its second-personal character tell us something important about the concept and nature moral obligation, it may also have genuine normative bite. It may help us see what it is "we owe to each other" and why we have normative reason so to act.

References

Bond, E. J. (1983). *Reason and Value*. Cambridge: Cambridge University Press.

Dancy, J. (2000). *Practical Reality*. Oxford: Oxford University Press.

Darwall, S. (1983). *Impartial Reason*. Ithaca, NY: Cornell University Press.

Darwall, S. (2002). *Welfare and Rational Care*. Princeton, NJ: Princeton University Press.

Darwall, S. (2003). "Desires, Reasons, and Causes," *Philosophy and Phenomenological Research* 67: 436–443.

Darwall, S. (2006). The Second-Person Standpoint: Morality, Respect, and Accountability. *Cambridge, MA: Harvard University Press.*

Darwall, S. (2011). "Demystifying Promises," in *Promises and Agreements: Philosophical Essays*, ed. Hanoch Sheinman. Oxford: Oxford University Press, pp. 255–276.

Darwall, S. (2012). "Bipolar Obligation," in *Oxford Studies in Metaethics*, v. vii, ed. Russ Shafer-Landau. Oxford: Oxford University Press, pp. 333–367. Also in Darwall, *Morality, Authority, and Law: Essays in Second-Personal Ethics I.*

Darwall, S. (2013a). Morality, Authority, and Law: Essays in Second-Personal Ethics I. *Oxford: Oxford University Press.*

Darwall, S. (2013b). Honor, History, and Relationship: Essays in Second-Personal Ethics II. *Oxford: Oxford University Press.*

Davidson, D. (1963). "Actions, Reasons, and Causes," *The Journal of Philosophy* 60: 685–700.

Enoch, D. (2011). "Giving Practical Reasons," *Philosophers' Imprint* 11: 4.

Foot, P. (1972). "Morality as a System of Hypothetical Imperatives," *The Philosophical Review* 81: 305–316.

Nagel, T. (1970). *The Possibility of Altruism*. Oxford: Clarendon Press.

McDowell, J. (1979). "Virtue and Reason," *Monist* 62: 331–350.

Moore, G. E. (1993). *Principia Ethica*, rev. ed. with the preface to the (projected) second ed. and other papers, ed. with an intro. Thomas Baldwin. Cambridge: Cambridge University Press.

Parfit, D. (1997). "Reasons and Motivation," *Aristotelian Supplementary Volume* 71: 99–130.

Parfit, D. (2011). *On What Matters*, 2 vols. Oxford: Oxford University Press.

Pettit, P., and M. Smith (1990). "Backgrounding Desire," *The Philosophical Review* 99: 565–592.

Raz, J. (1972). "Voluntary Obligations and Normative Powers, II." *Proceedings of the Aristotelian Society*. Supp. vol 47: 79–102.

Raz, J. (1986). *The Morality of Freedom*. Oxford: Clarendon Press.

Scanlon, T. M. (1982). "Contractualism and Utilitarianism," in *Utilitarianism and Beyond*, eds. Bernard Williams and Amartya Sen. Cambridge: Cambridge University Press.

Scanlon, T. M. (1998). *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.

Schapiro, T. (2009). "The Nature of Inclination," *Ethics* 119: 229–256.

Strawson, P. F. (1968). "Freedom and Resentment," in *Studies in the Philosophy of Thought and Action*. London: Oxford University Press.

Wallace, R. J. (1994). *Responsibility and the Moral Sentiments*. Cambridge, MA: Harvard University Press.

Watson, G. (1987). "Responsibility and the Limits of Evil: Variations on a Strawsonian Theme," in *Responsibility, Character, and the Emotions: New Essays in Moral Psychology*, ed. F. D. Schoeman. Cambridge: Cambridge University Press.

Wolf, S. (2012). *Meaning in Life and Why It Matters*. Princeton, NJ: Princeton University Press.

Part 5: The philosophy of practical reason as the theory of practical rationality

31. The normativity of rationality[5]

[5] Thanks to Kurt Sylvan for very helpful comments on a previous draft.

Errol Lord

Requirements are a dime a dozen. There are requirements of morality, requirements of parliamentary order, requirements of prudence, requirements of the local Boy Scout troop, requirements of prescriptivist grammar, requirements of rationality. Not all of these requirements are created equal. One important distinction to be made amongst requirements is between requirements that have *normative force* and those that don't. Some of the requirements just mentioned seem to have normative importance that others lack. The requirements of the local Boy Scout troop or of prescriptivist grammar don't seem to have normative force in the way that the requirements of morality and prudence do. The former requirements get a grip on you, if at all, only when you stand in particular relationships with them – when you are a member of the local Boy Scout troop, say. Morality and prudence, it is very commonly thought, have a kind of normative power all by themselves. Whatever relationship you have with them, their requirements get a grip on you.

One set of requirements that have historically been seen to be on stable normative footing are the requirements of *rationality*. Indeed, if anything, the requirements of rationality have been seen to be on the *steadiest* normative footing. This is evidenced by the fact that debates about the normative standing of morality and prudence have almost always used rationality as an anchor. Optimists about the normative force of prudence and morality usually aim to show that those requirements are normative *because* we are rationally required to comply with them.[1] Pessimists argue they lack normative force *because* we are not always rationally required to comply with them.[2]

These historical facts make it all the more surprising that it has become popular to think that rationality does not necessarily have normative force. Further, as we'll see, there are powerful arguments for this pessimistic view if you hold plausible views of normative force and popular views about the nature of rationality.

In the next section, I will explicate the main views about the nature of rationality. I will then flesh out the relevant conception of normative force. Section 2 will show that pessimism is very plausible given those views about normative force and the nature of rationality. The final two sections of the chapter will explore two ways out of the problem. The penultimate section will consider views that deny the relevant views about normative force. I will argue that these views have heavy burdens. In the final section, I will explore the sort of view that I've defended, which rejects the pessimist's conception of rationality in a way that opens the door for optimism about the normativity of rationality.

[1] See, for example, Korsgaard (1996) and Smith (1994).

[2] See, for example, Harman (2000) and Foot (1972).

1 Coherence and rationality

The thought that there is an intimate connection between rationality and coherence is ubiquitous. Given this, the recent literature on practical reason holds that the paradigm requirements of rationality explain why one is irrational when one's mental states are incoherent.

Here are some examples of the kind of irrational incoherence in question. It seems irrational to intend some end φ, believe that in order to φ one must intend to ψ, and lack an intention to ψ. Similarly, it seems irrational to believe p and believe $\neg p$ at the same time. Finally, it seems irrational to believe that one ought to φ and not intend to φ at the same time. While it is common ground to think that one is irrational when one is incoherent in these ways, it is controversial *which* requirements explain why one is irrational when one is incoherent. Two views are at the center of this debate. According to the first, one is irrational when one is means-end incoherent because one violates Means-End N when one is incoherent (cf. Schroeder (2009), Lord (2014b, 2011), Bedke (2009)):

> Means-End N: If A intends to φ and believes that in order to φ she must intend to ψ, then rationality requires A to intend to ψ.

The requirement operator only scopes over the consequent of the conditional – it takes narrow-scope over just the consequent. For this reason, those who accept requirements like Means-End N have come to be known as *narrow-scopers*. Narrow-scopers hold that one is rationally required to intend to ψ when one intends to φ and believes that one must intend to ψ in order to φ. Thus, they hold that you are forbidden from being means-end incoherent because you are required to intend to ψ when you have the antecedent attitudes (i.e., the attitudes mentioned in the antecedent of the conditional).

Narrow-scopers hold that a particular reaction is required when you are means-end incoherent; in the means-end case, one is required to intend to take the means. According to the second central view, one is merely required to *avoid* incoherence. This is because the second view holds that one is irrational when one is means-end incoherent because one violates Means-End W (cf. Broome (1999, 2002, 2013), Brunero (2010, 2012):

> Means-End W: A is rationally required to [intend to ψ if one intends to φ and believes that she must intend to ψ in order to φ].

The requirement operator scopes over the whole conditional in Means-End W – it takes wide-scope over the whole conditional. For this reason, those who accept requirements like Means-End W are called *wide-scopers*. Wide-scopers hold that all that rationality requires is that one not be means-end incoherent. This is because Means-End W does not require that one have any particular attitudes. It merely requires that one *not* be in the incoherent state.

The wide-scope view has become the dominant view. This is largely due to the arguments of Broome (1999), which aim to show that *if* the narrow-scope view is true, then the requirements of rationality cannot be *strict*. In order to be strict, requirements need to flat-out forbid things – it needs to be that one is *wrong* for not heeding the requirement's call. If narrow-scope requirements have this feature, then it looks like one can bootstrap into rational requirements to do bizarre things. To see this, imagine that I intend to become the King of Russia and believe that in order to become the King of Russia, I must intend to move to Siberia. If Means-End N is true, then rationality requires me to intend to move to Siberia. And if this requirement is strict, then it follows I am flat-out forbidden from not intending to move to Siberia.

This looks like a crazy prediction. The clearly irrational intention to become the King of Russia and the unhinged belief that moving to Siberia is a necessary means to becoming King should not have the rational power to require me to intend to move to Siberia. Fortunately for the wide-scoper, Means-End W does not make this prediction. This is because Means-End W *merely* forbids the incoherent combination. Thus, the wide-scope view does not predict that I am rationally required to intend to move to Siberia. It just predicts that I am forbidden from being in the incoherent state. So I can comply with Means-End W by doing the sensible thing and revising my antecedent attitudes. Thus, it looks like wide-scope requirements like Means-End W can have strict force without issuing bad predictions. Since rational requirements are plausibly strict – they are real *requirements*, after all – this is a virtue of the wide-scope requirements.

So goes a common and powerful argument for the wide-scope view. I will call it the Bootstrapping Argument, for it seeks to vindicate the wide-scope view by showing that only it can avoid bootstrapping into certain rational requirements by having irrational antecedent attitudes. The Bootstrapping Argument is the main reason for the ascent of the wide-scope view in recent literature about practical reason.

Before moving onto the nature of normative force, let me note an important wrinkle. As I understand it, the Bootstrapping Argument is *not* about normative force.[1] Rather, it is about the internal structure of rationality. It maintains that *even if rationality lacks normative force*, the wide-scope view is to be preferred because it doesn't allow for bootstrapping. To make this seem less strange, think of debates in normative ethics about the internal structure of morality. Those debates can seemingly be carried out

[1] Historical accuracy demands that I note that Broome's (1999) argument did invoke normative force. This is because he was linking strictness and normative force. This is not necessary, though, and the wide-scoper should be reluctant to do so. It is not clear what Broome's current view is. He does not appeal to bootstrapping in his arguments for wide-scoping in Broome (2013, ch. 7).

without settling the normative force of morality. Consequentialists and Kantians can debate the structure of morality without taking a stand about whether morality has normative force. As I see it, the debate about bootstrapping is like this.[2]

[2] Hussain (MS) is very explicit that he conceives of the argument in this way. Brunero (2010, 2012) is less explicit but also seems to think of it in these terms.

2 Normative force and normative reasons

Now that we know how many recent theorists have thought about rationality, we can turn to normative force. As in the last section, the goal is not to establish a particular view; rather, the goal is to explicate how normative force has been generally understood in the recent literature. Most discussions start (and end) with the idea that normative force is intimately connected to *normative reasons*. Broome (2007a), for example, writes 'a requirement on you to F is normative if and only if it constitutes a reason for you to F' (3). With this sort of view as a backdrop, Kolodny (2005) seeks to understand the relationship between the requirements of rationality and the requirements of reasons. At least at first, Kolodny thinks that is very plausible that the requirements of rationality have normative force. He also takes it that most will think it has normative force in virtue of its connection to the requirements of normative reasons. As we will see in a moment, he then argues that in general there are not conclusive normative reasons to do what rationality requires.[1]

With this rough characterization in hand, we can state some more precise views. Broome's idea previously is that a requirement to φ is normative just in case it – the requirement itself – *constitutes* a reason to φ. There are two important observations to make about this view. First, Broome holds that all requirements are grounded in *sources*. Legal requirements are paradigmatic; they are grounded in legislatures. Broome holds that each set of requirements has a source analogous to legislatures. He also holds that, at least for rationality, we can determine what the source requires independently of determining what there is normative reason to do. This is why (again, at least in the case of rationality) we can sensibly ask whether the requirements of rationality have normative force.

The second observation worth highlighting is that there are at least two ways to further flesh out Broome's idea.[2] According to a stronger version of the idea, rationality has normative force because whenever rationality requires one to φ, the fact that rationality requires one to φ provides a *decisive reason* to φ. When one has decisive reason to φ, one ought to φ, full stop. Call this version of the idea Strong Broomean

[1] Even after he gives his arguments for this conclusion, he tries to save the normative force of rational requirements. In the end, he holds that the requirements of rationality are only apparently normative (see Kolodny, 2005, 512–513 for a summary).

[2] Broome himself does this in Broome (2013, pg. 192–193).

Normative Force. According to a weaker version of the idea, rationality has normative force because whenever rationality requires one to φ, the fact that rationality requires one to φ provides at least some normative reason to φ. This weaker version of the idea doesn't take a stand about whether the reasons to be rational are always (or even usually) decisive. Let's call this Weak Broomean Normative Force.

Broomean Normative Force is not the only way of understanding the relationship between the requirements of rationality and normative reasons. It is anchored in Broome's legislative model of normativity. This view is not compulsory, nor is it particularly popular in its most general form. It is also restrictive. We can ask less restrictive questions about the normative force of rationality. In particular, we can ask whether there are always normative reasons or decisive normative reasons to do what rationality requires. These are less restrictive questions because they do not just ask whether the fact that rationality requires one to φ provides a normative reason to φ. I will say that rationality has Weak Normative Force just in case there are always normative reasons to do what rationality requires. Rationality has Strong Normative Force just in case there are always decisive normative reasons to do what rationality requires. At least as a starting point, then, we can hold that rationality has normative force just in case it has Weak Normative Force or Strong Normative Force. The question that has interested philosophers the most is whether rationality has Strong Normative Force. This is because it seems so plausible that we ought to be rational. But, again at least at first, it looks like we ought to be rational only if rationality has Strong Normative Force.

3 The normative impotence of coherence requirements

Now the stage is fully set for the most influential arguments for skepticism about the normative force of rationality. In the second and third subsections, I will sketch the two arguments that have received the most attention. Both arguments originate in Kolodny (2005). Before we get to those arguments, though, I will consider the question of whether the narrow-scope coherence requirements have normative force. As we will see, it is close to common ground that they do not. This plays a prominent role in Kolodny's arguments.

3.1 The normative impotence of narrow-scope coherence requirements

Suppose, again, that I intend to become the King of Russia and believe that in order to become the King of Russia, I must intend to move to Siberia. If Means-End N is true, then it follows that I am rationally required to intend to move to Siberia. Now I am, we can suppose, not in any actual position to become the King of Russia. Further, suppose that my belief that in order to become the King of Russia, I must intend to move to Siberia is based on a conspiracy theory involving proper lines of succession, Siberia, and my long-dead ancestors. My antecedent attitudes in this case are very much irrational.

If the narrow-scope means-end principle were true *and* rationality had Strong Normative Force, it would follow that I have decisive reason to intend to move to Siberia. This is obviously false. Thus, the requirements of rationality very plausibly do not have Strong Normative Force if they are all narrow-scope coherence requirements.[1]

What about Weak Normative Force? If the narrow-scope means-end principle were true and rationality had Weak Normative Force, then I would have some normative reason to intend to move to Siberia. It is compatible with this that the reason to intend to move to Siberia is a very weak reason. It might be that the balance of

[1] As Kurt Sylvan pointed out to me, this claim is only obvious if it is restricted to narrow-scope requirements that govern individual incoherent patterns (like means-end coherence). This reasoning isn't obviously good when it comes to more global narrow-scope requirements – for example, a narrow-scope requirement that takes into account all of one's mental states. See Brunero (2010) for discussion

reasons overwhelmingly supports staying in New Jersey. Given this, it is not easy to have reliable intuitions about the claim that the narrow-scope requirements have Weak Normative Force, for our intuitions about bare reason ascriptions are often unreliable (see Schroeder 2007 for discussion). Mark Schroeder uses this fact to defend the Weak Normative Force of the narrow-scope requirements in Schroeder (2004, 2005).[2]

Even if this defense is successful, though, it only wins the narrow-scoper a small battle. This is because Weak Normative Force doesn't establish what most want established, which is the claim that there is something *wrong* with irrationality – that we ought not be irrational. The Weak Normative Force of the narrow-scope rational requirements does not establish this. My having weak normative reason to intend to move to Siberia does not guarantee that I am doing something wrong if I intend to become King of Russia, believe that in order to become King I must intend to move to Siberia, and fail to intend to move to Siberia.[3] Something stronger must obtain in order to secure this. The most plausible candidate at this point is that rationality has Strong Normative Force.

As one might expect, the preceding argument is often used to further motivate the wide-scope view. For the wide-scope view does not obviously face the problem.[4] In the same way the wide-scope requirements avoid bootstrapping, they also avoid bad predictions about normative force. The wide-scope requirements might have Strong Normative Force even though I ought not intend to move to Siberia. For it only follows from the Strong Normative Force of the wide-scope means-end requirement that I ought not be incoherent. Since I can become coherent by dropping my end or means-end belief, I can do what the wide-scope requirement requires without forming the intention to move to Siberia. Since Broome (1999), this has been seen as one of the chief virtues of the wide-scope account.[5]

This is the backdrop to Kolodny's (2005) influential arguments. I will now turn to those.

---

of this sort of idea.

[2] For another sort of defense, see Bratman (2009); it should be noted that Schroeder later changed his mind on this point in Schroeder (2009).

[3] Schroeder gives up the strategy pursued in Schroeder (2004, 2005) in Schroeder (2009) for this reason.

[4] Some have argued that in some cases wide-scope requirements do give rise to this problem. See Greenspan (1975), Setiya (2007), Schroeder (2009), Bratman (2009) for discussion.

[5] Indeed, many take this to be *the* bootstrapping argument against the narrow-scope view (see, e.g., Kolodny 2005). This is largely because Broome (1999) originally provides a bootstrapping argument against the narrow-scope view by arguing that only the wide-scope view can vindicate the normative force of rationality. However, as I read things, the primary conclusion to draw from Broome's arguments is about the *structure* of rationality. He appealed to claims about normative force because he *assumed* that rationality had normative force. We can run the bootstrapping argument without appeal to normative force just by pumping intuitions about what rationality itself requires.

3.2 The symmetry argument and its discontents

3.2.1 The argument

Kolodny's first argument is indirect in an important way. Rather than argue that the wide-scope requirements themselves are normatively impotent, he argues that at least some requirements of rationality are in fact narrow-scope.[1] Given the arguments in the last subsection, it plausibly follows that at least the narrow-scope part of rationality is normatively impotent. Given this indirectness, the main action in the first argument is about the structure of rationality itself.

Kolodny's argument against the wide-scope requirements turns on a principle connecting the requirements of rationality to *reasoning*. Some rational requirements, claims Kolodny, are about *processes* rather than just *states* of an agent's mind. So-called state requirements are just about which combinations are permitted or forbidden (usually at a single time). Process requirements, on the other hand, govern transitions between states. They tell us which transitions are permitted or forbidden.

Kolodny's question is whether the process requirements are wide-scope. He argues that they are not by appealing to what he calls *the reasoning test*. The reasoning test says, roughly, that if one is rationally required to (revise attitude A or revise attitude B), then one can 'rationally resolve' the conflict between A and B by revising either A or B. In order to rationally resolve the conflict, one must *reason* one's way out of it. As Kolodny is thinking of things, when we reason out of a conflict, we do so by reflecting on the *content* of the states that are in conflict. We notice in some way, that is, that the content of one of the states demands the revision of some other state and we revise the former state because of this awareness.

With this test in hand, Kolodny argues that some conflicts are governed by narrow-scope rather than wide-scope requirements. We can make his point by focusing on means-end coherence even though he uses different examples.[2] So suppose I intend to get a coffee, believe that in order to get a coffee I have to intend to go inside, but fail to intend to go inside. Means-End W entails that I am rationally required to (drop my

[1] Although this is the only claim Kolodny needs for the argument, he in fact holds the stronger view that all the requirements are narrow-scope.

[2] He focuses on conflicts between beliefs about what one has reason to believe/intend and a lack of the belief/intention. What is crucial is that the relevant conflicts involve the absence of an attitude.

intention to get a coffee *or* drop my belief that in order to get a coffee I must intend to go inside *or* form the intention to go inside). In order for Means-End W to pass Kolodny's test, it must be possible for me to reason out of the conflict by moving from my lack of intention to go inside to the dropping of one of my antecedent attitudes. But it is not clear that I can do this.

In fact, given Kolodny's way of understanding reasoning, this is *impossible*. For on Kolodny's understanding, rational revision always involves moving from the content of some attitude to the formation or dropping of a different attitude. But my lack of intention has no content. Even if you don't have this strong a view of rational revision, it is not clear that I can rationally resolve this conflict by dropping an antecedent attitude *in light of* my lack of an intention to go inside. If this is right, then Means-End W fails Kolodny's reasoning test. If the reasoning test is a true test of process requirements, then Kolodny has shown that Means-End W is false. Means-End N, on the other hand, passes the test; it is possible for me to rationally resolve the conflict by reflecting on the contents of the antecedent attitudes and forming the intention to go inside in light of my awareness of those contents.

If sound, Kolodny's argument (i) rules Means-End W out and (ii) provides strong support for Means-End N. But given the argument in the last subsection, Means-End N lacks normative force. Thus, at least one rational requirement lacks normative force.

3.2.2 Criticisms of the symmetry argument

There are three moving parts to this argument. First, there is the reasoning test. Second, there is the claim that there are process requirements. Third, there is the claim that narrow-scope requirements are normatively impotent. All three of these claims have been questioned.

Way (2011) argues that the wide-scoper should reject the reasoning test. His reason is that the test demands too much. Too see this, imagine that rather than merely believing that in order to get coffee, I must intend to go inside, I *know* this. Thus, in order for Means-End W to pass the reasoning test, it would have to be possible for me to rationally revise a belief that constitutes knowledge by reflecting on lack of an intention to go inside. But this seems too strong. This is reason to doubt the reasoning test.[3]

Broome (2007b) questions the second moving part by questioning whether there are any narrow-scope process requirements that (i) could be plausibly accepted by the narrow-scoper and (ii) could support Kolodny's skeptical argument. He takes Kolodny committed to the idea that, since processes take time, the attitudes mentioned in the antecedent of the requirement will be tokened at an earlier time than the attitudes mentioned in the consequent. So the narrow-scope process requirements will require

[3] For another reason, see §4 of Broome (2007b).

an attitude – an intention, say – because one has an attitude – a belief, say – at some earlier time. But Broome thinks this is independently implausible. For we can imagine a case where one has the belief and intention at the earlier time but rationally drops both at the later time (because of new information, say). Broome acknowledges that Kolodny might have other processes in mind that don't fall prey to this objection. In that case, though, Broome is doubtful that the resulting process requirements will give rise to the features that Kolodny's skeptical argument relies on.[4]

I will leave discussion of Kolodny's third moving part – that the narrow-scope requirements are normatively impotent – for the next section.

Before moving on, it should be noted that the basic idea behind Kolodny's argument has been explored in other ways by other philosophers. Notice that at bottom, Kolodny's thought is that there is a kind of *asymmetry* between various parts of the conflicting sets of attitudes. For Kolodny, this asymmetry is cashed out in terms of rational resolution of the conflict. Others have argued against wide-scope requirements by appealing to different asymmetries.

Schroeder (2009) argues in favor of narrow-scope requirements by appealing to the thought that wide-scope requirements sanction *rationalization*. For example, the wide-scope requirement forbidding akrasia seems to hold that dropping one's belief that one ought to φ in light of the fact that one isn't going to φ is rationally on a par with forming an intention to φ in light of one's belief that one ought to φ.[5] Lord (2014b) (see also Schroeder 2004, 2015) argues for the narrow-scope view by pointing out that, for example, Means-End W maintains that you comply with the instrumental requirement either by revising your antecedent attitudes or by intending the means. Means-End N, on the other hand, maintains that you *escape* from the instrumental requirement by revising your antecedent attitudes whereas you comply by intending the means. There are reasons to want to mark the difference between escaping and complying. Given the background assumption that the narrow-scope requirements are normatively impotent, these arguments, if sound, show that at least some requirements of rationality are normatively impotent.

3.3 The missing reasons argument

The Symmetry Argument in its various forms has been a source of great controversy. The other argument Kolodny gives has been much more widely accepted.[6] Let's suppose that the Symmetry Argument does not work and that all of the requirements of rationality are wide-scope. We can directly test the normative force of the wide-scope requirements by asking which facts provide reasons to do what the wide-scope

[4] See Kolodny (2007b) for a reply to Broome.

[5] For pushback see Way (2011), Broome (2013).

[6] Broome also presents an influential version of this second argument in Broome (2005a, 2005b, 2008, 2013).

requirements demand. The Missing Reasons Argument contends that we will not find any reasons to do what the wide-scope requirements demand.

Let's focus once again on instrumental rationality and on the case from the last subsection – I intend to get some coffee, believe that in order to get coffee I must intend to go inside, but fail to intend to go inside. Initially, one might be puzzled by the claim that I have no reason to do what Means-End W demands. After all, it's very plausible that I do have reason not to be the way I am. Given the setup of the case, I have reason to intend to go inside, and if I do, then I will be complying with Means-End W. So, one might think, I do have reason to do what Means-End W demands in this case.[7]

As tempting as this looks, it is generally assumed that showing that there is always reason to react in some way that leads to compliance with a wide-scope requirement is *not* enough to show that there are reasons to do what the wide-scope requirement demands. The wide-scope requirements forbid incoherence directly. They do not ensure coherence by requiring that one have particular attitudes; rather, they ensure coherence by directly forbidding incoherence.

Thus, it has been assumed that, if there are reasons to do what the wide-scope requirements demand, they are reasons that *directly* speak in favor of being coherent. My reason to intend to go inside is not like this. It doesn't speak in favor of coherence. It speaks in favor of intending to go inside. My being instrumentally coherent is just a side effect of reacting in the way the reason speaks in favor of.[8]

Now that we are clear about what needs to be shown, the problem becomes acute. Which facts *directly* speak in favor of coherence? In our example, there are reasons for me to intend to get coffee, reasons for me to believe that I must intend to go inside to get coffee, and reasons for me to intend to go inside. Which facts go over and above this and speak in favor of (not intending to get coffee or not believing I must go inside to get coffee or intending to go inside)? Kolodny considers some options.[9] First, it might be that we always have instrumental reasons to be coherent – it might be that by being coherent, we are always doing something else we have reason to do. While it's plausible that we always do something else we have reason to do by being coherent in *certain ways*, it is not plausible that we always do something else we have reason

[7] This kind of argument can seemingly be given whenever I am incoherent, for it seems plausible that there will always be some way out of the incoherence that I have reason to take. In the King of Russia case, for example, the way out supported by reasons is dropping the antecedent attitudes. See Lord (2014a, 2017a), Kolodny (2007a) for more discussion.

[8] There are good reasons not to go in for the thought that the reason to intend to go inside is also a reason to be coherent. The argument I just gave tacitly relies on the principle that if r is a reason to φ and by φ-ing I will ψ, then r is a reason to ψ. This principle is dubious. Ross's (1944) paradox brings this out. The fact that my friend needs to know how I'm doing is a reason to post the letter. By posting the letter, I am (posting the letter or burning the letter). It doesn't seem to follow that the fact that my friend needs to know how I am is a reason to (post the letter or burn the letter). For more on this, see Lord (2017b, 2018).

[9] Cf. Broome (2008, 2013), Way (2010b), Lord (2018).

to do by being coherent in *any* way. In the King of Russia case, I do something I have reason to do by dropping my intention to become King of Russia, but it is far from clear I do something I have reason to do by intending to move to Siberia. Even if I did, this reason wouldn't be decisive and thus this won't vindicate the Strong Normative Force of Means-End W.

The second option is that the fact that we must in general be coherent in order to count as agents at all provides a reason to be coherent. Even if we grant that our agency depends on being coherent much of the time, it seems that all that follows is a reason to be coherent *some of the time*. I won't cease to be an agent by failing to intend to go inside. So it's not clear why this fact about agency provides me a reason to be coherent. And, once again, even it did, it doesn't seem like this reason would be very strong. In the King of Russia case, I am going very badly wrong if I intend to move to Siberia even if my agency depends on my not being coherent.

The final option is that coherence itself is a final good along things like knowledge, friendship, pleasure, wisdom, and so on. Most have found this suggestion unintuitive. Kolodny sums the point up well when he says that it is 'outlandish that the kind of psychic tidiness that [a rational requirement] enjoins should be set alongside such final ends as pleasure, friendship and knowledge' (2007a, 241). Further, just like the other options, even if we did think coherence was a final good, the reason provided by this goodness would not always be decisive. Thus, this looks unlikely to vindicate the Strong Normative Force of the wide-scope requirements.

Many have been convinced by the Missing Reasons argument that rationality is normatively impotent. The argument does seem to undermine the Strong Normative Force of the wide-scope requirements. As noted previously, Weak Normative Force is harder to debunk given the unreliability of intuitions about bare reason ascriptions. Still, many are also persuaded by these arguments that establishing Weak Normative Force is a losing battle.

4 Responding to the challenge

While many have been moved to skepticism by the foregoing arguments, there has been serious resistance. In this section, I will sketch two popular ways of resisting. One way denies the views of normative force used to generate the challenge, while the other denies that rationality is tied to coherence in the way assumed by the skeptic.

4.1 Alternative views of normative force

The most popular way of responding to skepticism has been to deny that normative force is spelled out in terms of Strong or Weak Normative Force. One option that some have flirted with is to insist that there is more than one source of normative force. Reasons provide one source, whereas rationality provides another. As far as I know, no account like this has been spelled out in detail.[1] For this reason, I will only mention it and move on.

The most worked-out alternative account of normative force relies on a distinction between *objective* and *subjective* normative force.[2] The skeptical arguments tacitly appealed to the claim that Strong and Weak Normative Force only involve *objective* normative reasons. These are the reasons that are provided by the *facts*. Rationality has Strong Normative Force just in case there are always decisive *objective* normative reasons to do what rationality requires.

A common thought is that we don't have decisive *objective* reasons to be rational, but we do have decisive *subjective* reasons to be rational. Subjective reasons are not tied to the facts the way objective reasons are; rather, they are tied to our (often misguided) *perspective* on the world. So, for example, according to Schroeder (2009), p is (roughly) a (decisive) subjective reason for A to φ just in case A believes p and p would be an (decisive) objective reason to φ if p were true.

To illustrate the appeal of this, consider an epistemic example – what I'll call closure. It's plausible that p and if p, q provide decisive objective reason to believe q when they are true. Thus, if Schroeder's account of subjective reasons is on the right track, I have decisive subjective reasons to believe q whenever I believe p, and if p, then q. There is thus always something subjectively wrong with believing p, if p, then q, but failing to believe q.

[1] For sketches, see Southwood (2008), Hussain (MS), Langlois (2014). For critical discussion, see Levy (FC).

[2] See Schroeder (2009), Way (2012, 2010a), Parfit (2011), Whiting (2014).

The first problem to note is that there are technical difficulties with fully generalizing this to all incoherent combinations. The instrumental case highlights the problem. In order for the view to account for the instrumental case in the way it accounts for the closure case, it has to be that whenever one intends to φ, one has decisive subjective reason to intend to φ. If this is true, then one can explain what's subjectively wrong with means-end incoherence by appealing to Schroeder's account of subjective reasons and the claim that if one has decisive objective reason to φ and decisive objective reason to believe that in order to φ one must intend to ψ, then one has decisive objective reason to intend to ψ.

The problem is that it is far from clear that whenever one intends to φ, one has decisive subjective reasons to φ. In Schroeder's view, in order for this to be true, one needs to have some beliefs the content of which would provide decisive objective reason to φ were they true. But why think it's necessarily true that every time one intends to φ one will have such beliefs? Partly motivated by this problem, Way (2010a) proposes that, necessarily, the fact that intending to ψ is necessary for φ-ing is itself a decisive *wide-scope* objective reason to (intend to ψ or not intend to φ). If this is right (and a Schroederian account of subjective reasons is right), then whenever one believes that intending to ψ is necessary for φ-ing, one will have decisive subjective reason to (intend to ψ or not intend to φ). Complying with this demand ensures instrumental coherence so Way's account explains why something is always going subjectively wrong when one is instrumentally incoherent.

The second problem to note is that while the machinery provides some helpful explanatory resources, it is not clear why subjective reasons provide the right kind of normative force. In Schroeder's view, in the King of Russia case, I have decisive subjective reasons to intend to move to Siberia. The obvious question to ask in this context is: So what? Why think that this verdict has any normative significance?[3] Why think this has any interesting bearing on which attitudes are the attitudes to have?

To be fair to Schroeder, he does have a general account of normativity, according to which what it is to be normative is to be analyzed in terms of objective normative reasons (see Schroeder 2007). If this view is true, then Schroeder's narrow-scope requirements are normative since they are analyzed in terms of subjective reasons which are in turn analyzed in terms of objective reasons. While this is something, it's not clear it even begins to answer the skepticism.

For Schroeder is clearly using 'normative' in a different way than I have been. He is using 'normative' in a very wide sense. Purely evaluative notions (e.g., goodness) are normative, in Schroeder's understanding.[4] I, on the other hand, have been talking about

[3] Way's account of the instrumental case has an advantage here since he just holds that I am rationally required to (intend to move to Siberia or not intend to be King of Russia). It's worth noting, though, that Way is a narrow-scoper about other combinations like akrasia (cf. Way 2013). So he will be open to this sort of skeptical question about those requirements.

[4] This is made clear by the fact that he thinks that goodness is normative in virtue of being analyzed in terms of objective reasons.

a narrower notion of normativity. This notion is only tied to the *deontic*. Requirements are normative in my sense only if they are tied to what you ought to do in some sense. So, even if Schroeder's narrow-scope requirements are normative in his wide sense, it's not clear they are normative in my deontic sense.

4.2 Alternative views of rationality

Another way of resisting the original challenge is to deny the views of rationality that give rise to it. This has been pursued most clearly in Lord (2014a, 2017a, 2018) and Kiesewetter (2018). According to my preferred view, rational requirements are determined by the objective normative reasons that one *possesses*. The reasons that one possesses are the reasons that are within one's ken in a special way. There are several different accounts of possession in the literature (see Lord 2018, Whiting 2014, Sylvan 2015, Schroeder 2008, Williamson 2000, Neta 2008, Gibbons 2013). According to my account, what it is to possess a reason r to φ is to be in a position to manifest knowledge about how to φ for r. I argue further that in order to be in a position to manifest knowledge about how to φ for r, one must be in a position to know r (see Lord 2018, chs. 3–4 for discussion and defense).

Notice that if this view of rationality is true, it immediately follows that rationality has Weak Normative Force. This is because in order to be rationally required to φ, according to this view, you have to possess objective reasons to φ. Thus, there will always be objective reasons to do what rationality requires according to this view. Securing Weak Normative Force is thus trivial for this view.

Strong Normative Force is a different matter. This is because we can fail to possess all of the reasons. Thus, there are cases where the reasons you possess decisively support φ-ing even though when all of the objective reasons are weighed, some alternative to φ-ing is decisively supported. Thus, if what you ought to do full stop is determined by all of the objective reasons, this account of rationality cannot vindicate the Strong Normative Force of rationality.

Unsurprisingly, many have denied that what you ought to do full stop is determined by all of the reasons. Indeed, one of the main debates in the literature on ought is about this issue (see Jackson 1991, Graham 2010, Thomson 2008, Lord 2015, 2017b, 2018, Kiesewetter 2011). I have argued that it is independently plausible that what you ought to do full stop is determined by the normative reasons you possess. This is because it is plausible that in order for some reason r to be eligible to obligate you, you must be able to correctly respond to r. And in order to have this ability, I claim, you have to possess r (for my defense, see Lord 2015, 2017b, 2018; for critical discussion, see Way & Whiting FC).

If what you ought to do full stop is determined by the objective reasons you possess and what you are rationally required to do is determined by the objective reasons you possess, it follows that what you ought to do full stop just is what you are rationally

required to do. Thus rationality has Strong Normative Force. Indeed, its normative significance is even deeper. For it is consistent with Strong Normative Force that there are things one ought to do full stop that one is not rationally required to do. This is not so in my view. In my view, you ought to φ full stop just in case you are rationally required to φ. So rationality, in my view, has ultimate normative importance. The view's ability to explain this is a strong reason to adopt it and thus reject the views of rationality that motivated Kolodny's skepticism.

5 Conclusion

This chapter has carried out four main tasks. First, it introduced common views about the nature of rationality. These views tether rationality to facts about coherence. Second, it introduced common views about the nature of normative force. These views tether normative force to normative reasons. Third, it showed that when you combine the views about rationality with the views about normative force, it is very plausible that rationality lacks normative force. The fourth task was to explicate two ways to respond to this skepticism about the normative force of rationality. In the end, I argued that rationality does have normative force by rejecting coherentist views about rationality. The requirements of rationality are determined by the possessed normative reasons. What you ought to do is also determined by possessed normative reasons. Thus, the requirements of rationality have normative force.

References

Bedke, M. (2009). The Iffiest Oughts: A Guise of Reasons Account of End-Given Reasons and End-Given Oughts. *Ethics*, 119.

Bratman, M. (2009). Intention, Belief, Practical, Theoretical. In S. Robertson (Ed.), *Spheres of Reason*. Oxford: Oxford University Press.

Broome, J. (1999). Normative Requirements. *Ratio*, 12, 389–419.

Broome, J. (2002). Practical Reasoning. In J. Bermudez & A. Millar (Eds.), *Reason and Value: Themes from the Moral Philosophy of Joseph Raz*. Oxford: Oxford University Press.

Broome, J. (2005a). Does Rationality Give Us Reasons? *Philosophical Issues*, 15, 321–337.

Broome, J. (2005b). Have We Reason to Do as Rationality Requires? A Comment on Raz. *Journal of Ethics & Social Philosophy*, 1.

Broome, J. (2007a). Requirements. In J. J. Toni Rønnow-Rasmussen, Björn Petersson & D. Egonsson (Eds.), *Homage à Wlodek: Philosophical Papers Dedicated to Wlodek Rabinowicz*. Online Tribute (url: www.fil.lu.se/hommageawlodek/).

Broome, J. (2007b). Wide or Narrow Scope? *Mind*, 116(462), 359–370.

Broome, J. (2008). Is Rationality Normative? *Disputatio*, 11.

Broome, J. (2013). *Rationality Through Reasoning*. Oxford: Blackwell.

Brunero, J. (2010). The Scope of Rational Requirements. *The Philosophical Quarterly*, 60(238), 28–49.

Brunero, J. (2012). Instrumental Rationality, Symmetry, and Scope. *Philosophical Studies*, 157(1), 125–140.

Foot, P. (1972). Morality as a System of Hypothetical Imperatives. *The Philosophical Review*, 81(3), 305–316.

Gibbons, J. (2013). *The Norm of Belief*. New York: Oxford University Press.

Graham, P. A. (2010). In Defense of Objectivism about Moral Obligation. *Ethics*, 121(1), 88–115.

Greenspan, P. (1975). Conditional Oughts and Hypothetical Imperatives. *The Journal of Philosophy*, 72(10), 259–276.

Harman, G. (2000). Is There a Single True Morality? In *Explaining Value and Other Essays in Moral Philosophy*. Oxford: Oxford University Press.

Hussain, N. (MS). The Requirements of Rationality. Manuscript, Stanford University.

Jackson, F. (1991). Decision-Theoretic Consequentialism and the Nearest-Dearest Objection. *Ethics*, 101(3), 461–482.
Kiesewetter, B. (2011). "Ought" and the Perspective of the Agent. *Journal of Ethics & Social Philosophy* 5 (3): 1–24.
URL: WWW.JESP.ORG.
Kiesewetter, B. (2018). *The Normativity of Rationality*. Oxford: Oxford University Press.
Kolodny, N. (2005). Why Be Rational? *Mind*, 114(455), 509–563.
Kolodny, N. (2007a). How Does Coherence Matter? *Proceedings of the Aristotelian Society*, 107, 229–263.
Kolodny, N. (2007b). State or Process Requirements? *Mind*, 116(462), 371–385.
Korsgaard, C. (1996). *The Sources of Normativity*. Cambridge: Harvard University Press.
Langlois, D. (2014). *The Normativity of Structural Rationality*. PhD thesis, Harvard University, Cambridge.
Levy, Y. (FC). Does the Normative Question About Rationality Rest on a Mistake? *Synthese*.
Lord, E. (2011). Violating Requirements, Exiting from Requirements, and the Scope of Rationality. *The Philosophical Quarterly*, 61(243), 392–399.
Lord, E. (2014a). The Coherent and the Rational. *Analytic Philosophy*, 55(2), 151–175.
Lord, E. (2014b). The Real Symmetry Problem(s) for Wide-Scope Accounts of Rationality. *Philosophical Studies*, 170(3), 443–464.
Lord, E. (2015). Acting for the Right Reasons, Abilities, and Obligation. In R. Shafer-Landau (Ed.), *Oxford Studies in Metaethics*, vol. 10. Oxford: Oxford University Press, pp. 26–52.
Lord, E. (2017a). The Explanatory Problem for Cognitivism About Practical Reason. In D. W. Conor McHugh & Jonathan Way (Eds.), *Normativity: Epistemic and Practical*. Oxford: Oxford University Press. Manuscript, Princeton University Press.
Lord, E. (2017b). What You Ought to Do and What You're Rationally Required to Do (Are the Same Thing!). *Mind* 126 (504):1109–1154.
Lord, E. (2018). *The Importance of Being Rational*. Oxford: Oxford University Press.
Neta, R. (2008). What Evidence Do You Have? *British Journal for the Philosophy of Science*, 59(1), 89–119.
Parfit, D. (2011). *On What Matters*. Oxford: Oxford University Press.
Ross, A. (1944). Imperatives and Logic. *Philosophy of Science*, 11(1), 30–46.
Schroeder, M. (2004). The Scope of Instrumental Reason. *Philosophical Perspectives*, 18, 337–364.
Schroeder, M. (2005). Instrumental Mythology. *Journal of Ethics & Social Philosophy*, 1.
Schroeder, M. (2007). *Slaves of the Passions*. Oxford: Oxford University Press.
Schroeder, M. (2008). Having Reasons. *Philosophical Studies*, 139(1), 57–71.

Schroeder, M. (2009). Means-End Coherence, Stringency, and Subjective Reasons. *Philosophical Studies*, 143(2), 223–248.
Schroeder, M. (2015). Hypothetical Imperatives: Scope and Jurisdiction. In R. Johnson & M. Timmons (Eds.), *Reason, Value, and Respect*. Oxford: Oxford University Press.
Setiya, K. (2007). *Reasons Without Rationalism*. Princeton: Princeton University Press.
Smith, M. (1994). *The Moral Problem*. Oxford: Blackwell.
Southwood, N. (2008). Vindicating the Normativity of Rationality. *Ethics*, 118(1), 9–30.
Sylvan, K. (2015). What Apparent Reasons Appear to Be. *Philosophical Studies*, 172(3).
Thomson, J. J. (2008). *Normativity*. Open Court Press.
Way, J. (2010a). Defending the Wide-Scope Account of Instrumental Reason. *Philosophical Studies*, 147(2), 213–233.
Way, J. (2010b). The Normativity of Rationality. *Philosophy Compass*, 5(12), 1057–1068.
Way, J. (2011). The Symmetry of Rational Requirements. *Philosophical Studies*, 155(2).
Way, J. (2012). Explaining the Instrumental Principle. *Australasian Journal of Philosophy*, 90(3), 487–506.
Way, J. (2013). Intentions, Akrasia, and Mere Permissibility. *Organon F*, 20, 588–611.
Way, J. & Whiting, D. (FC). Perspectivism and the Argument from Guidance. *Ethical Theory and Moral Practice*, 1–14.
Whiting, D. (2014). Keep Things in Perspective: Reasons, Rationality, and the a Priori. *Journal of Ethics and Social Philosophy*, 8, 1–22.
Williamson, T. (2000). *Knowledge and Its Limits*. Oxford: Oxford University Press.

32. The eclipse of instrumental rationality

> [T]he ideal of usefulness permeating a society of craftsmen – like the ideal of comfort in a society of laborers or the ideal of acquisition ruling commercial societies – is actually no longer a matter of utility but of meaning. It is 'for the sake of' usefulness in general that *homo faber* judges and does everything in terms of 'in order to.' ... [But] utility established as meaning generates meaninglessness.
>
> –Arendt (1958: 154)

Introduction

Many philosophers of practical reason assume that practical rationality is partly constituted by the suitable coordination of means and ends – that is, by *instrumental rationality*.[1] This assumption is clear in much of the literature on requirements of rationality, in which means-end coherence remains a standard example of what rationality requires, alongside coherence between one's normative beliefs and one's intentions ('Enkrasia') and consistency between one's intentions. The assumption is also clear in much of the literature on practical reasoning: while many reject the instrumentalist view that practical reasoning is always of means and never of ends, many also grant that coordinating our means to our ends is a central case of practical reasoning.[2]

Instrumental rationality was argued to be a myth in Raz (2005). But the objections he raised to the normativity of instrumental rationality turned out to be special cases of broader worries about the normativity of coherence explored by Kolodny (2005, 2007) and Broome (2005, 2007).[3] Hence, it seems fair to say that most in the current literature see no *special* problem about instrumental rationality. Most agree that *if* there are any fundamental requirements of coherence, an instrumental principle is among them; those who follow in the footsteps of Raz – for example, Kiesewetter (2017) and Lord (2018) – are best understood as denying that there are fundamental requirements of coherence.

In this chapter, I want to question this lingering consensus. I think there are *special* problems about the normativity of instrumental rationality which don't merely reflect

[1] It was once common to hold that practical rationality is *wholly* constituted by instrumental rationality. See Gauthier (1987) and Dreier (1996) for examples, and Nozick (1993: 133) for an illustration of the perceived dominance of this view. The view is often pinned on Hume, but Sayre-McCord's contribution to this volume shows that this may be wrong. In the literature on rational requirements which grew out of the pioneering work of Broome (1999), it is more common to hold that there are both instrumental and non-instrumental requirements of rationality. Still, it is unusual for theorists to accept coherence requirements but deny that instrumental coherence is among them. For a thoroughly non-instrumental view, see Hampton (1998), and see Korsgaard (2009) for the view that the instrumental principle is merely an aspect of a categorical requirement of rationality.

[2] See, for example, Audi (2001), Broome (2013), Dancy (2018), Millgram (1997) and Vogler (2002) for figures who grant that there is non-instrumental reasoning but assume that instrumental reasoning remains a central case of practical reasoning. See Kolnai (1962) and Williams (1981) for the view that all genuinely practical reasoning is of ends. These figures occupy an interesting space in agreeing with cognitivists about practical reason that instrumental reasoning reduces to theoretical reasoning, while adding that it is not practical for this reason.

[3] For more discussion in this volume, see the Introduction and the contributions by Lord and Morton and Paul.

broader problems about the normativity of coherence requirements. But I think we needn't fret, since the patterns of reasoning that the instrumental principle allegedly underwrites shouldn't have been regarded as instrumental from the outset. Hence, I will argue that we can do without instrumental rationality. This eclipse of instrumental rationality is, I believe, good news for the unity of reason. I have argued elsewhere that epistemic rationality is wholly non-instrumental.[4] The story about practical rationality I give here contributes to a unified picture of the epistemic and practical.

It is worth emphasizing that I will defend this view while accepting that practical rationality has a significant *structural* component not reducible to either (i) the pressures of *apparent reasons* (*pace* Kiesewetter 2017, Kolodny 2005, and Lord 2018), (ii) requirements of theoretical rationality (*pace* cognitivists like Setiya 2007 and Wallace 2001) or (iii) to a categorical imperative (*pace* Hampton 1998 and Korsgaard 2009). I agree with Vogler (2002) that there is a fine-grained order to practical reasoning that is omitted in views which regard all practical reasoning as reasoning of ends. I merely deny that this order is a *calculative* order, as Anscombe (1957) said. In particular, in place of the calculative structure established by 'in order to' relations holding between intentions and acts, I substitute an order of meaning better captured by 'for the sake of' relations holding between intentions/acts and values.[5] I swap ends for values which are not merely 'to be promoted' and replace means-end relations with relations reflecting the internal structure of the values for the sake of which one acts. Hence instrumental structure is eclipsed by subjective axiological structure, with the latter understood in a non-consequentialist way.[6]

With these ideas in mind, here is the plan. I begin in §1 with some terminological clarifications and a more precise statement of my main claims, together with some disclaimers. I turn in §2 to give special reasons for skepticism about instrumental rationality. §3 shows that the practical phenomena commonly assumed to be underpinned by instrumental rationality can be better explained by non-instrumental structural rationality. §4 sketches a more specific non-instrumental account which better captures the order the instrumental principle was meant to capture. I conclude in §5 by showing how this picture fits nicely with a wider strategy for vindicating the normativity of rationality that I have developed elsewhere.

---

[4] See, for example, Sylvan (2014, 2018, 2020).

[5] The approach may also have been implicitly accepted by Continental philosophers who were critical of what they called 'instrumental reason' (especially Adorno and Horkheimer 1944/1979, Arendt 1958, Gorz 1989, Horkheimer 1947, 2012, Marcuse 1964, and Weber 1921/1968). The paper stalks these figures in its footnotes, and its title is an allusion to Horkheimer's *Eclipse of Reason* and *Critique of Instrumental Reason*.

[6] My understanding of subjective axiological structure also distinguishes my approach from *teleological* approaches as normally understood. Whether it differs from properly Aristotelian approaches is another matter. Note that a close translation of the Greek phrase which unpacks the idea of final cause – '*to hou heneka*' – is *that for the sake of which*. It is possible that Aristotle shouldn't be regarded as a teleologist if being a teleologist means explaining normativity in terms of *aims* which are fundamentally to be *promoted*; for more discussion, see Johnson (2005).

1 Instrumental rationality: some preliminaries and disclaimers

1.1. The face-value understanding of instrumental rationality and reasoning

What is instrumental rationality? I work with a face-value understanding that takes the word 'instrumental' in its ordinary sense, not as shorthand for an intuitively broader concept or a technical concept.[1] Hence, I assume that if a form of reasoning does not conclude with the reasoner's *intending to use something as a means* in any pretheoretically recognizable sense of 'means', we lack good reason to call it 'instrumental'. As we will see, there are many patterns of reasoning which don't conclude in such instrumental intentions that ought to be distinguished from instrumental reasoning. By correctly distinguishing these forms of reasoning from instrumental reasoning, the face-value understanding helps to carve practical reason at its joints.

To be more precise, the face-value understanding assumes that instrumental rationality is characteristically manifest in reasoning which moves toward the intention to use an apparent means to bring about an end, where the end is treated by the agent as having a value that is 'to be promoted' (i.e., to be brought about for its own sake).[2] The understanding hence takes the premise-attitudes of properly instrumental reasoning to be

> (i) intentions to bring about certain states of affairs (the agent's ends), and
>
> (ii) beliefs that performing certain actions or using certain resources (the agent's means) would help to bring about these states of affairs,

[1] Vogler (2002: 163) gives a broader characterization that elides distinctions I want to draw: 'The view I have been laying out isn't exactly that calculative or technical practical reasoning is entirely a matter of finding "causally efficacious means" by which to attain determinate ends. This is one form such reasoning might take, but it could just as well trace constitutive part-whole relations, or else involve straightforward demonstrative inference'. As we'll see, reasoning founded on constitutive relations is often fundamentally different from means-end reasoning. I agree with Audi (2001) that it is unhelpful to obscure this difference with the technical notion of a 'constitutive means'.

[2] Hence the ultimate ends of an agent's reasoning will appear from their perspective to have *final value* of the kind consequentialists use to explain rightness; see Pettit (1989) and Scanlon (1998: Ch.2) for this characterization of consequentialism and Scanlon for an argument that not all fundamental value is final value in this sense.

and it takes the conclusion-attitudes to be

> (iii) *instrumental intentions* to perform certain actions *in order to* bring about the states of affairs targeted by the premise-attitudes.

An example would be reasoning from (i) the intention to bring about peace and (ii) the belief that disarmament would help to bring about peace to (iii) the intention to pursue disarmament in order to promote peace. Here the conclusion intention apparently has a kind of structural rationality relative to the premise attitudes. If there is such a thing as instrumental rationality, this is a paradigm case.

1.2 Some contrasting phenomena

Surprisingly enough, the face-value understanding is more fine grained than many in the literature and helps to contrast instrumental rationality with several phenomena often conflated with it.[3]

To begin to see why, note that the instrumental intentions with which instrumental reasoning ends are a special case of *derivative pro-attitudes*. If we are interested in carving at the joints of practical reasoning, we should pay attention to forms of reasoning which end in other kinds of derivative pro-attitudes. In parallel to the distinction between non-instrumental value and fundamental value (which aligns with the more familiar distinction between instrumental and derivative value),[4] we should allow for valuing which is non-instrumental but derivative. Reasoning guided by such valuing is not 'calculative'.

To see non-instrumental but derivative pro-attitudes in action, consider an example inspired by Korsgaard (1983). I assign special value to the scarf you gave me, even though I rarely wear it and I have other scarves that keep me warmer. My valuing is not instrumental. It is derivative, however: I don't treat the scarf as having *ultimate* value. Instead, I value the scarf *because* it is a token of your friendship. It is just that the 'because' here signals no instrumental relation. To use a different example which applies a model from Hurka (2001), I value art-appreciation, and my valuing of it is not instrumental. But this valuing is also not bedrock: I value art-appreciation *because* I value art and I think appreciation is the fitting response to art.

Intentions are pro-attitudes that can manifest non-instrumental ways of valuing. Hence they can be derivatively yet non-instrumentally rational in the same way. This point matters, because it suggests that processes of reasoning that have been *modeled*

[3] Instrumental reasoning is often merely portrayed (i) as starting with the belief that Y-ing is *necessary* for X-ing rather than the more specific belief that Y-ing is *a necessary means* to X-ing, and (ii) as ending with an intention to Y which is not qualified as an *instrumental intention*. See Brunero (2020) for a striking illustration of this tendency.

[4] See Korsgaard (1983) for one example and Sylvan (2014, 2018, 2020) for discussions of the importance of this distinction for understanding the value of epistemic rationality.

instrumentally might well be better understood as concluding in different kinds of derivatively rational intentions.

Another contrast that the face-value understanding supports is between instrumental reasoning and what I'll call *constitutive* reasoning (which Millgram 2001 called *specificationist* reasoning). Suppose I think I should respect your privacy. I think about what would be involved in doing this now. I decide not to enter your room without knocking and hearing you say it is OK to enter. Here it would misrepresent me to regard *promoting respect for privacy* as an end that I have and to regard knocking and waiting as a means to bringing about this end. I just intend to respect your privacy, and I intend to knock because that is what it is to respect your privacy on this occasion.[5] To be sure, I will bring about the state of affairs in which your privacy is respected. Hence we can *model* my action as a 'constitutive means' to bringing about this state of affairs. But we would not correctly describe *my reasoning* if we portrayed it in this way.

If this is right, we should not think that patterns of reasoning such as the following necessarily involve instrumental rationality:

I intend to X

I believe that Y-ing would constitute X-ing in this case

So I intend to Y

If the background belief which leads me to the intention to Y is the belief that I ought to respect your privacy, then it would be a misrepresentation to portray me as thinking that privacy is an end to be promoted, with my action understood as a means to bringing about this state of affairs. For there must be a form of reasoning that enables one to properly respond to values to be respected by determining what respect consists in on the occasion and then leading me intend to do that thing.

This is not yet to pass judgment on consequentialism or the consequentializing project[6] or to stack the deck in favor of non-consequentialism.[7] My minimal suggestion at this stage is that it is possible to think like a non-consequentialist and to reach intentions by reasoning that embodies non-consequentialist ways of valuing. This is a modest claim. It is consistent with this claim that such reasoning doesn't track the objective norms and that I shouldn't reason in this way. It is also consistent with this claim that this way of thinking is right because it promotes the good. All I say so far is this: (i) it is possible to treat something as a value to be respected and not (merely) to be promoted, and (ii) there is a form of reasoning from more general to more specific intentions which embodies such valuing.[8]

[5] See Hurley (2018: 32) for the same point.

[6] See Dreier (1993) and Portmore (2007) for defenses, and Schroeder (2007b) for a critique.

[7] As Hurley (2018) emphasizes, it is only to *avoid* stacking the deck *in favor of* consequentialism.

[8] I agree Raz (2011, 2016) that *no* values are fundamentally to be promoted. But this view isn't required for the view defended here. Only briefly in section 2] will I mention how to piece together this Razian view with the myth view of instrumental rationality. (Note that Raz didn't explicitly connect these ideas and still allowed for instrumental value in Raz (2005).)

Even if X is an end to be promoted, the move to a more specific intention won't be instrumental if it only involves specification. Consider a kind of example from Richardson (1994: 77). I want to order something light and vegetarian. I see that there is only one option (that salad). I think: 'Actually, that would be very nice.' I form the intention to have it. Here I won't be eating the salad as a means to the more general end of eating something light and vegetarian. I want to eat it for its own sake. Still, I concluded that I will eat it on the basis of practical reasoning that moved from a more general to a more specific intention. If I merely needed to eat *something* vegetarian and the salad seemed tolerable, perhaps we could imagine that I order it as a means to eating something vegetarian. But my reasoning is not always correctly portrayed in this way. There are many permissible ways of transitioning from an intention/belief pair to a further intention that are not instrumental.

1.3 Face-value instrumental reasoning and rationality in more detail

With those contrasts made, let's consider a fuller statement of the face-value understanding:

Instrumental reasoning is reasoning from an end to-be-promoted and the belief that X-ing is a means to promoting that end to the intention to do X at least partly for the reason that it would help to promote the end. Ends-to-be-promoted are naturally embodied in intentions. Other motivational pro-attitudes within our rational control could embody ends-to-be-promoted (e.g., desires). What is essential is that the reasoning ends with an at least partly *instrumental pro-attitude*. It cannot end with an intention to X just for the sake of X-ing. It also cannot end with an intention to X for the sake of Y, where Y is not understood as a value to be promoted. It also cannot end with an intention to do an action that includes an instrument (a piano) but constitutes a larger intrinsically valuable activity (playing beautiful piano music). It must end with an intention to use a means to bring about an end.

We can then distinguish between (i) sufficient means, which are X-ings that will alone produce the end; (ii) partial means, which are X-ings that will *help* to produce the end; and (iii) necessary means, which are essential steps in the process of producing the end-state. (iii), it is worth noting, excludes *preconditions*, since they are undertaken *before* the process of producing the end-state starts. Eating breakfast is not part of writing a paper in the afternoon, though it may be a causally necessary precondition. The distinction here seems worthwhile. Preconditional actions are distinct from means in the same way that enabling conditions (e.g., oxygen) are distinct from causes (e.g., fire). To take another example, when I say that I am going to the park to fly a kite, I don't regard going to the park as a *means* of flying the kite. I regard it as *putting me in a position* to do so.

An *instrumental requirement of rationality* will then be any 'iffy' principle that says that you are rationally required to have a certain instrumental intention *if* and *because* you have a certain end-to-be-promoted and a certain instrumental belief. A *pressure of instrumental rationality* will be any *apparent reason* to have an instrumental intention generated by the appearance that certain instrumental facts hold and one's having an end-to-be-promoted.[9]

Although it is not normally stated in the literature on requirements of rationality, it is crucial to add the 'and because' clause to a candidate principle of instrumental rationality. We must allow that there might be *other* reasons you could be rationally required to have an instrumental intention given certain other mental states on some occasion, and these reasons *might not support belief in any instrumental requirement.* Here we should compare principles of rationality with other explanatory normative principles. Compare an unexplanatory principle which says that it is right to do X if C with an explanatory principle which says that it is right to do X if C *because* C. It is, for example, right to be nice to the people next to you if you're on a plane. But it is not right to be nice to them *because* you're on the plane.

Finally, I leave open whether instrumental requirements are to be formulated in a wide-scope way or a narrow-scope way.[10] I just assume that coming to have an instrumental intention on the basis of an end-to-be-promoted and an instrumental belief which coheres with that instrumental intention is what counts as complying with the alleged instrumental requirement.

1.4 Claims and disclaimers

With the foregoing clarifications in the background, I can now state the two main claims I oppose:

> *The Status Explanation Claim*: In cases that can be *modeled* instrumentally, the fact that the conclusion-intentions are rational is *explained* by the fact that they comply with requirements or pressures of instrumental rationality.
>
> *The Necessary Glue Claim*: Rational practical reasoning, intending and acting are necessarily held together, at least in significant part, by instrumental reasoning and responsiveness to apparent instrumental pressures.[11]

In opposition, I will argue that (i) apparent instrumental relations don't do the explanatory work that they are commonly assumed to do, and that (ii) we needn't

[9] I endorse Fogal's (forthcoming) claim that rationality involves responding to pressures, not just requirements.

[10] For discussion, see Errol Lord's contribution to this volume.

[11] Few explicitly make these claims; the one clear example is Vogler's (2002) insightful defense of

fret, since we don't need them to confer a sufficiently fine-grained order on thought and action. Along the way, I will defend some contrasting positive claims:

> *The Valuing Claim*: In the cases that some model instrumentally, the rationality of the conclusion-intention is better explained by the fact that it manifests a derivative but non-instrumental way of valuing (constituents of) the intended event.
>
> *The Non-Instrumental Order Claim*: Rational practical reasoning, intending, and acting exhibit a fine-grained but non-instrumental order. This order is grounded in for-the-sake-of relations linking one's acts/attitudes to one's non-instrumental values.

These claims distinguish my rejection of instrumental rationality from other approaches that dispense with fundamental instrumental coherence requirements, such as reasons-first approaches (Kolodny 2005, Raz 2005, Kiesewetter 2017, Lord 2018), cognitivist approaches (Setiya 2007, Wallace 2001) and the most familiar Kantian approaches (Korsgaard 2009, Hampton 1998). Indeed, I see this chapter as one installment in a wider rejection of instrumental ideology that would also target instrumental *value* and *reasons*, which are not targeted by many of these other theorists.

Still, although a wider anti-instrumentalist agenda is in the background, this chapter is only explicitly about rationality in what Scanlon (1998) called 'the narrow sense'. Hence, as a final disclaimer, I note that here I'm not directly opposing views about reasons like Schroeder's (2007a) Humeanism or Portmore's (2011) consequentialism. Some of my arguments might generalize to these views. But some would be more questionable: an analogue of the argument in §2.2 may, for example, seem too quick against Humeanism and consequentialism, for the reasons in Schroeder (2007a: Ch.2) and Railton (1984). But alienation matters more clearly for the theory of rationality.[12]

instrumental reason. But I assume that the fundamental requirements of rationality are justified by the fact that they *explain* the rational status of certain attitudes or combinations of attitudes. Hence, anyone who takes the instrumental principle to be a fundamental principle of rationality implicitly accepts the first claim. It is hard to see why one would believe the first claim unless one believed the second. But for a defense of the second, see Anscombeans like Vogler (2002) and Schwenkler (2019), who take it to capture Anscombe's suggestion that action exhibits a 'calculative order'.

[12] The importance of alienation for the theory of action was noted in Lavin (2013). But I defend the near opposite of Lavin's claim. Lavin suggested that if there were *basic* actions, they would necessarily involve alienation. I will be arguing that if there were *instrumental* actions, they would necessarily involve alienation.

2 Against the Status Explanation Claim

I will now turn to give several arguments against the Status Explanation Claim (henceforth 'SEC').

2.1 Argument from the explanation of canonical examples

The first argument involves looking at the kinds of examples that might seem best modeled instrumentally and then maintaining that the rationality of the derivative intentions that hold these cases together is not instrumental on the face-value understanding.

I take it that if instrumental rationality is going to be on display anywhere, it will be in an extended course of action which can be divided into steps or phases, where the course leads to completion. If it is worth its salt, instrumental rationality should be leading us from one step to the next, on to the literal end. The trouble is that the steps in many courses of action do not seem best described as *means* to the completion of the action. Instead, they seem to be *parts* of the action-in-progress. The intentional doing of any of the steps at a given time *just is* the intentional doing of the action, albeit incomplete at the time, and in the process of completion. Consider building a toy house out of blocks for fun. I am not laying the blocks as a means to building the house. Laying the blocks is part of building the house: here I am *already* performing some of the action which is my end.[1]

It seems better to understand this case as one in which I intend a whole non-derivatively and then intend the parts derivatively, because their unity *is* the whole intended as an end. While the intentions to do each smaller step are derivatively rational, and rational relative to the end, the relation of derivation isn't a paradigmatic instrumental relation.

[1] For a longer defense of this picture, see Thompson (2008). Curiously, Thompson (p. 89) says that he is interested in ''instrumental' or 'teleological'' rationalization, and that his model of naïve rationalization offers a better picture of this. As I will suggest subsequently, it may be fine to describe some rationality as *teleological*, but this should be sharply distinguished from anything *instrumental*, as well as from anything 'calculative'.

If this is all correct, then insofar as we can attribute the status of derivative rationality to my intention to lay this block, this status will not be *best explained* by an instrumental requirement or by instrumental pressures. Instead, it seems better to think that it makes sense for me to intend to lay this block because that is *part of what is involved in* building the house. Of course, not all the things I do which are intelligible in light of my desire to build the house are parts of building the house. But those other things aren't means either: they are *preconditions*. Buying the blocks, or clearing space for them, for example, are not in themselves *means* to building the house.

If we work with the face-value understanding of instrumental rationality, then, it doesn't seem that the cases where instrumental rationality would be most likely to be exhibited are cases where it is needed to do any explanatory work. For we can divide our smaller activities into two groups: (i) the activities which lead up to one's undertaking some project and (ii) the activities which are parts of the progress of the project. The activities that most straightforwardly come to mind under each heading do not seem properly described as means to the completion of the activity. They are either preconditions or parts. Parts can be *modeled* from the outside as constitutive means. But the intentions which hold together a complex action are not best understood as instrumental. The thinking behind them can be wholly non-instrumental.[2]

2.2 Argument from alienation

Having an instrumental intention which targets one's own actions is, I believe, anomalous on closer inspection. I am not clear that it is possible to wittingly sustain such an intention, at least for long. For, as I'll argue, having such an intention would involve a sort of *alienation* that structural rationality should frown upon, not require.[3] And it is worth stressing in advance that the familiar consequentialist tool for circumventing alienation from Railton (1984) won't help here: for here we are not dealing in the first instance with objective norms but rather subjective principles.

[2] Audi (2001: 84) puts the point well in a discussion of pleasure: '[W]hat we intrinsically want *for* pleasure is not properly said to be wanted as a means to pleasure. Wanting something for pleasure is wanting it for the (presumed) intrinsic qualities of it that make it attractive to one *as* pleasurable; it is not wanting it as a causal or other contingent producer of pleasure. To want something for pleasure is to want it in the anticipation of pleasure *in* realizing it'.

[3] The alienation at issue is a generalization of what Marx (1988) discusses in the *Economic and Philosophical Manuscripts*. As Benhabib (1994) usefully summarizes, a central concern of the Frankfurt school was to generalize Marx's points about alienation to life under capitalism in general to yield (using Horkheimer's phrase) a 'critique of instrumental reason'; see especially Horkheimer (1947, 2012), Adorno and Horkheimer (1944) and Marcuse (1964). As I see it, this project was incomplete and never lived up to Horkheimer's label. It focused only on instrumental reason *under capitalism* instead of giving a critique of instrumental reason in general, which is exhibited in non-capitalist social structures and the solitary life of Robinson Crusoe (a favorite example of neo-classical economists).

There are independent lines of Continental thought which get closer to targeting instrumental

Before I give the argument, a word of caution is in order about the kind of example I will use and my strategy in using it. I will consider cases in which instrumental motivation seems *most transparently* to be on display. Instrumental motivation may be less transparent in other cases if it is present. But I think that is because there will also be non-instrumental motivations working alongside instrumental motivations in other cases. What I assume is that if there is such a thing as instrumental rationality, it is possible for it to be fully manifest in the most transparent cases of instrumental motivation. I will then suggest that these cases are cases of alienation and that instrumental motivation is the source of the alienation. I will also assume that genuine rational requirements should not be such that complying with them transparently *constitutes* alienation. Of course, one might be inclined to say in my examples that there is more going on in the psychology of the agent which contributes to their alienation. But I will want to say that the crux of the alienation is instrumental motivation.

Let me proceed. The literature on requirements of rationality has gotten us used to the idea that our *actions* can be means. A standard statement of the most often discussed instrumental principle, after all, has the simple form:

> *IP-Simple*: It is a rational requirement that if you intend to A and believe that B-ing is a necessary means for A-ing, then you intend to B, where A-ing and B-ing are actions.[4]

After reading those words many times, they can sound like they pick out something familiar. But I think we must take a step back and reflect on how strange it would be to transparently conceive of one's own actions as *means* and to intend these actions *because* they are means.

I allow that we are familiar with means in ordinary life, but the means with which we are familiar are not actions: they are *mere things*, such as forks and paintbrushes. We use these things, and our usings are actions, and these actions perhaps have 'instrumental value'.[5] But it doesn't follow from these claims that we regard our *actions*

---

reason as such. As Ridley (2018) sees it, this project may have been in the background of Nietzsche's central contributions to the philosophy of action. Arendt (1958: Part IV]) ought to be the *locus classicus*. Weber (1921/1968), Heidegger (1954/1977) and Gorz (1989) also get closer than the Frankfurt school to a wider critique of instrumental reason.

[4] See Bratman (1987), Hill (1973), Kiesewetter (2017), Korsgaard (1997), Scanlon (2007), Schroeder (2004) and Way (2012) for examples of this way of thinking about the means covered by the instrumental principle. Brunero (2012), Harman (1976), Kolodny and Brunero (2018) and Lord (2018) give formulations that instead cast one's *intentions* as means; the arguments in this section apply even more strongly, I think, to this formulation. Owing to background views about the nature of intention, Broome (1999, 2013: 157) gives a more unusual formulation in which what is intended is an *event*, which represents a different approach discussed in Section 3] . At least in translation, Kant (1785/2012) varies between the first and last formulations.

[5] I assume for the moment that there is instrumental value. But see Prichard (1937/2002: 214) for doubts about the category of instrumental value which complement the view here. (Thanks to Jonathan Dancy for drawing this to my attention.)

as means and intend them for that reason. Instead, we normally will these actions as parts of some larger undertaking. One's use of the paintbrush is part of one's activity of painting, for example.

A vertigo creeps in when we start to conceive of our smaller activities as means that don't share in the value of the project to which they were meant to be contributions. Conceiving of them in this way detaches them from the larger meaningful activity in which they were formerly installed. It may detach them from our agency. If I could think of possible movements of my hands as tools I could use to achieve some end, and I exploited these movements with that objective in mind, the movements might rightly seem like puppetry. Partly for this reason, it is hard to get myself to conceive of my acts in this way.

I can think of cases in which this feeling is more familiar. But they fail to involve full rationality, at least if rationality is something of value. And it is plausible that what *makes* them fall short is the fact that they involve treating one's actions as means. Some examples might be cases in which I have to *get myself* to do something in which I see no value, but which has some chance of producing some unspecified later advantage, or cases in which I'm merely working on the basis of incentives (e.g., factory labor). In such cases, I can get myself to do the required act only by manipulating myself in some way, as in cases in which I try to respond to pragmatic reasons for belief. Yet it also seems these are the kinds of cases in which I am most transparently responding to instrumental pressures by treating my action as a means. When I can fit the action into some larger meaningful project, it is more appropriate to think of my action as part of something larger which I intend as an end. It gets a share of the same value that this larger thing has as an end.

2.3 Technical knowledge doesn't mark out a distinct field of practical rationality

One might think §2.2 only shows that the instrumental principle should never have been stated in a way that portrays *actions* as the relevant means. A different option is to push the means back into the world, where they belong. This option better captures the Baconian idea that technical knowledge makes us masters of nature, not parts of the machinery through which nature is controlled. But this view leads to a different problem for SEC.

The problem centers around the fact that it is unclear why there would be a distinctive sub-compartment of practical rationality devoted to the use of tools. A type of rationality should not be marked off from others just because it covers actions which involve certain kinds of objects (tools) or certain kinds of dealings with objects (treatment as tools). We can subsume subjectively appropriate forms of object-treatment under a more general principle equally applicable to non-instrumental cases. This broader

principle is the subjective analogue of the principle that there is reason of the right kind to respond to X in the way that is fitting to X (e.g., to desire the desirable, to esteem the estimable, to envy the enviable):

> *The Fitting Treatment Principle*: If you believe that X is fittingly treated in way W, then there is subjective reason of the right kind to intend to treat X in way W.

This principle captures our understanding of how to treat tools as a special case of our understanding of how to treat any objects of action. If I conceive of something as a toaster, I conceive of its function as being to toast. Hence it is sensible for me to use it according to its function by putting bread in it. But there is no new or distinctive kind of rationality in play here. The rationality in play is the same in play when I envy the apparently enviable or esteem the apparently estimable.

Conceiving of something as an instrument involves conceiving of it as being fittingly treated in a certain way. And so the belief that something is an instrument can be sensibly heeded by acting in certain ways. But the rationality which is on display in such cases is not distinctive. Servicing the apparently serviceable is a special case of X-ing the apparently X-able. Indeed, some cases of using instruments may involve no instrumental intentions: playing piano is most often done for its own sake, though it obviously involves using the piano.

Perhaps one could claim that there is a distinctive kind of knowledge – technical knowledge – invoked by the practical reasoning that others have deemed instrumental. But this fact still doesn't give rise to a distinctive form of practical rationality. This point was made before by Kolnai (1962), Wiggins (1975) and Williams (1981). Kolnai (1962: 187) put it especially well:

> So far as the physician confines himself to the determination of suitable curative means ... he does not deliberate but performs the theoretical activities of recalling to his mind relevant knowledge, looking up textbooks for more information, considering the peculiarities of the case in hand, weighing probabilities, comparing the average efficacy of various methods in similar cases, and so forth. He does what a consulting physician, not responsible for any decision, might to just as well for him. The knowledge he brings to his practical task is ampler and more exact but not of a logically distinct nature than my wholly unpractical knowledge.

To be sure, knowledge of the instrumental properties of objects may play an enabling role in helping one to acquire know-how. And know-how is distinctively practical. But properly understood, these points just put us on the other horn of my overarching dilemma. When a person manifests knowledge how to A by B-ing, her B-ing constitutes her A-ing and is known to do so in virtue of her know-how: her B-ing is a *way* of A-ing

in a constitutive sense, not a means. In this case, the agent does not treat her B-ing as a means to A-ing: the only thing treated as a means is the object.

Knowledge of the instrumental properties of the object is not on its own sufficient for practical knowledge, as Kolnai noted. Even if intellectualism about know-how were true, it is not *this* propositional knowledge which guides one's action: it is knowledge of a *way* of acting that is presented to one as potentially constituting one's intentional A-ing.[6] Technical propositional knowledge might play an *enabling* role in allowing me to grasp that way of acting, but once I grasp it, it is my direct apprehension of the way that grounds my knowledge of how to open the door.

We can now combine the points in this section and the last to get a larger argument against SEC:

1. We can either take the means of means-end coherence to be (a) *actions* (the implicit view in the literature), or (b) *mere things* which are treated in a certain way through one's acting.

2. If (a), the instrumental principle is false (and hence SEC is false).

3. If (b), SEC is false: the rationality of one's use of the thing is better explained by a combination of fittingness and the constitutive rationality of action guided by know-how.

4. Hence, SEC is false.

2.4 Arguments from subsumption and embeddedness

I turn to a fourth argument. It is similar in spirit to the first but consistent with a larger concession to instrumental thinking. This argument involves looking at a kind of rational activity that seems to have face-value instrumental structure but then arguing that the deeper explanation of one's rationality in these cases is non-instrumental. The suggestion will be that the apparently instrumental relations in play have significance only as special cases of a more general non-instrumental relation.

Let's start with the phenomenon. It seems clear that we often do one thing *in order to* do another. I go to the park in order to feed the ducks, for example. Relatedly, it seems that an intention to X can rationally link to an intention to Y via a practical basing relation expressed by 'in order to'. This basing-relation is at least a *teleological* relation, where Y-ing is the *aim* and X-ing is part of the process of fulfilling the aim. But

[6] It is telling that different stripes of intellectualists like Stanley and Williamson (2001) and Bengson and Moffett (2012) were inclined to appeal to ways rather than means, so that canonical intellectualism reduces knowledge of how to A to knowledge of a way W that W is a way to A, rather reducing it to knowledge of a means M that it is a means to producing some state of affairs.

it also appears to be instrumental. To be sure, I want to explain away this appearance, and suggested a recipe for doing so earlier. Still, once we confront plausible activity descriptions featuring the 'in order to' construction, we may be less inclined to explain it away. We might be more inclined to say that our initial face-value understanding was too narrow.

Suppose we concede that 'in order to' expresses a real instrumental basing relation. Does it follow that we must accept SEC? No, for two reasons. The first is that SEC gives a *fundamental* story about why certain intentions are rational. We have conceded that instrumental basing relations may be part of the superficial story. But we are not forced to include them in the fundamental story. For the 'in order to' relation is plausibly a special case of a more general relation, and there might be good reason to prefer a fundamental explanation featuring this more general relation.

Let's consider the relationship between the two constructions from the Arendt epigraph: 'in order to' and 'for the sake of'. It is clear that 'X-es for the sake of Y' does not entail 'X-es in order to Y', because 'for the sake of' can relate an action to (a) non-actions (e.g., persons for whose sake we act, or values for the sake of which we act) and (b) actions which are not fundamentally cases of bringing about states-of-affairs (e.g., respecting the law). Hence, we should not try to reduce 'for the sake of' facts to 'in order to' facts. To do so, as Arendt (1958: Part IV) emphasizes, is to over-extend instrumental reasoning. But we *can* reduce 'in order to' facts to 'for the sake of' facts. Consider 'He went to fridge to get some milk'. We can translate this sentence into the 'sake' ideology as follows: 'He went to the fridge for the sake of getting some milk'. This second sentence is not as elegant. But it is not false, ungrammatical or nonsensical. It is true if the first is true and *vice versa*.

These facts suggest that the 'sake' locution is *more general* than the 'in order to' locution. But there are good reasons to want more fundamental explanations of rational status to invoke more general ideology. Compare physics. There are different kinds of physical forces – for example, *contact forces* such as frictional forces and *non-contact* forces such as gravitational force. Suppose we want to explain why some object of fixed mass accelerated. We could give an ordinary explanation invoking a specific kind of contact force. It would normally be more elegant to do so: 'The ball moved because I kicked it' is crisper than 'The ball accelerated because I applied a force to it'. But the fundamental explanation will go via a force law that doesn't discriminate between contact and non-contact forces. Hence we shouldn't invoke contact forces specifically if we want the most fundamental explanation of why the ball accelerated. If we are interested in the fundamental laws of rationality, we have a similar reason to appeal to the most general features of actions that do the work. Doing so illuminates deeper similarities between otherwise different-seeming cases, in the way that Newton's laws illuminate the similarities between the motions of balls and planets.

We should not assume that 'in order to' explanations are fundamental. They can be subsumed under 'for the sake of' explanations in ways that reveal an underlying

similarity with cases only captured by 'for the sake of' explanations. If so, we should not accept SEC even if we agree that 'in order to' explanations are face-value instrumental.

This is not all we can say to put 'in order to' in its place. A second point is that 'in order to' explanations are tolerable only given the background assumption that the agent to whom they apply has some irreducibly 'for the sake of' values. This point is hard to see only because charity requires us to trust that a background story is available unless we have overriding reason to treat the agent as irrational.

Here the 'sake'-based translation is illuminating. Suppose I go to the fridge for the sake of getting some milk. Unless I am mad, this cannot be the *only* thing I can say about what I am doing. For suppose I ask myself why I am going to the fridge for the sake of getting milk, and there is literally no wider story I can conjure. If I am not mad, I will feel lost and wonder what I'm doing. If that doesn't happen, something's gone wrong: I have values exhibiting the kind of irrationality that Parfit (1984) noted with the beloved example of 'Future Tuesday Indifference', so that getting milk from the fridge – whether for nourishment or otherwise! – just has bedrock desirability for me.

One might try to respond by saying that we can always appeal to other values which are not irreducibly 'for the sake of' values. But here Arendt had an insight. In the chapter from which the epigraph is drawn, she suggests that any attempt to terminate the regress which appeals *only* to values-to-be-promoted will only kick the can down the instrumental road. Stopping with anything conceived of by the agent as merely 'to be promoted' in the way that nourishment is 'to be promoted' will either lead to alienation or reveal Future-Tuesday values. To be sure, it might initially appear satisfying to invoke, say, happiness. But that is only because happiness is also a value *for the sake of which* we act, not *just* something that we strive *to produce* or *get*. Conceiving of it in that way puts us on the hedonic treadmill. It would be worse than Future-Tuesday-indifference to have jogging on the hedonic treadmill as one's ultimate end. So, in saying that we want happiness for its own sake, we are *not* saying we want to have it in order to have it.

At the very least, this conclusion should seem compelling if 'in order to' is taken at face-value as expressing an instrumental relation. If 'in order to' means 'as a means to', it is irreflexive. Ends are not things which are means to themselves. Ends stand outside of the chain of in-order-to relations and can only terminate the regress of purposive action by doing so. Perhaps for this reason, the longer phrase Aristotle used to explain the notion of a *telos* (*to hou heneka*) is properly translated as 'that for the sake of which', not 'that end as a means to which'.[7] If 'in order to' has a reading closer to 'to hou heneka', then the relevant meaning will not be purely instrumental.[8]

[7] See Johnson (2005).

[8] In discussing whether the themes of this chapter are compatible with his (2019) reading of Anscombe, John Schwenkler pointed out to me that the most general phrase needed to understand the structure of action is not the one expressed by 'in order to A' but rather simply by 'to A', and he observed that the latter is far from being clearly instrumental. He hypothesized that the core relation

2.5 Argument from the value of rationality

My arguments so far have been internal to the theory of practical reason. I have been arguing that careful reflection on the structure of practical rationality doesn't support SEC and hence that we lack sufficient reason to believe that the instrumental principle is a fundamental principle of rationality. There is also an external argument we can give if we assume that rationality matters. It parallels some arguments I have given before for rejecting attempts to explain epistemic rationality instrumentally in Sylvan (2014, 2018, 2020).

The argument rests on the following assumptions about the significance of rationality:

> *The Necessary Value Claim*: Necessarily, if a mental state manifests rationality, that fact as such makes that mental state *pro tanto* better than it would otherwise be.
>
> *Derivativeness*: Although necessarily possessed by a mental state, the value which inheres in a mental state in virtue of being rational is *derivative*, relative to a more fundamental value (it is just that this *derivative* value may be *non-instrumental*).

These are claims about the value *simpliciter* of rationality; in other work, I defended the idea that epistemic rationality necessarily has a special, non-instrumental kind of derivative *epistemic* value relative to the more fundamental *epistemic* value of truth. But it is also possible to give a narrower argument from the assumption that rational *action* necessarily has a certain kind of intrinsic value *for the agent*: namely, it makes the agent's action *meaningful* for her. Indeed, I think these arguments are related, since meaningfulness is not only good for the agent but good *simpliciter*.

Given these assumptions, if we want to explain the value that some instance of rationality in a particular mental state or action has, we need to appeal to some relation R that the state or action bears to fundamental value such that:

> *Constraint on R*: Necessarily, if a mental state bears R to a more fundamental value, this fact as such makes that mental state better than it would be if it didn't bear R to V.

In previous work on epistemic rationality, I argued that R cannot be an instrumental relation to promoting accuracy (or any other plausible fundamental epistemic values), partly by extending points from Jones (1997) and Zagzebski (2000). Both suggested that the fact that justification necessarily makes a belief better is incompatible with

---

is one of *directionality* rather than *instrumentality*. This hypothesis would render his picture of the structure of action compatible with the view that non-instrumental forms of intentional directedness might underwrite action rather than ones properly labeled 'calculative'.

justification having value *merely as a product of an instrumentally valuable process*, where the underlying non-instrumental value is accuracy. Zagzebski made this point through her famous coffeemaker analogy: if a cup of coffee is already good, the fact that it was produced by reliable coffeemaker does not make it any better. Inverting a thought from Carter and Jarvis (2012), I added in Sylvan (2018) that the fact that a bad cup of coffee was produced by a reliable coffeemaker also doesn't make it any better, and I suggested that this was a general point about products of instrumentally valuable processes. It is not, however, a general point about *derivative value*, since there are non-instrumental forms that escape this reasoning. Hence, I suggested that R must be a non-instrumental relation, such as the relation of *being a fitting response* to some more fundamental value. I added to this point in Sylvan (2020) by suggesting that we need a non-instrumental model to explain why rational beliefs have epistemic value even in worlds where rational belief-forming processes are not truth conducive (e.g., skeptical scenarios).

What I didn't appreciate before is that this point casts more general doubt on SEC and instrumental rationality. If complying with the instrumental principle were to ground some measure of rationality, this rationality would have to have necessary but derivative value. Yet it seems clear that complying with the instrumental principle *as such* could only have instrumental value. If there are cases in which instrumentally rational intentions seem to have some further sort of value, that is not *just* because these intentions comply with the instrumental principle. But if rationality as such confers some necessary value on a mental state, then it would also seem to follow that instrumental relations cannot alone make a mental state rational.

Note that it is unhelpful to respond by insisting that instrumental rationality necessarily has instrumental value. Perhaps one could argue that instrumentally rational belief-forming processes as a type necessarily have instrumental value, though it is not plausible in skeptical scenarios that they will have *real* rather than *merely expected* instrumental value. But the token property of being instrumentally rational does not itself necessarily have instrumental value. It won't have such value if one's end isn't achieved. If one's end is achieved, there is no longer any good in having the mental state.

Hence, the 'swamping problem' for instrumental explanations of epistemic value seems to extend to instrumental explanations of practical rationality. This fact was noted in passing by Arendt (1958: 154–155): 'an end, once it is attained, ceases to be an end and loses its capacity to guide and justify the choice of means'.[9] And as Arendt was mainly observing, this point reveals that instrumental relations don't ground a kind of *rationality* at all. For whatever rationality is, it is necessarily something of intrinsic but derivative value.

[9] At this stage of her argument, she used this point to defend the other side of Prichard's thought that alleged instrumental value is not a kind of value: alleged value as an end-in-itself is also not a kind of value. Granting Prichard's point, her conclusion is a corollary: for ends are the subordinate business partners of means.

3 Against the Necessary Glue Claim

I turn to a briefer discussion of the Necessary Glue Claim, since much of the work needed to appreciate its falsity has already been done.

It is worth noting first that this claim is weaker than SEC. SEC attempts to limn the *grounds* of the practical rationality of certain patterns of reasoning. The Necessary Glue Claim merely holds that apparent instrumental relations and transitions *necessarily hold together* these patterns of reasoning, which is consistent with their rationality having some deeper non-instrumental explanation. Hence, not all good arguments against SEC yield good arguments against the Necessary Glue Claim.

To refute the Necessary Glue Claim, it is enough to show that there are *available* rational patterns of reasoning not held together by apparent instrumental relations which would lead us to all the proper conclusions to which instrumental reasoning would lead us. It is not necessary to show that we never employ instrumental reasoning. Perhaps the Frankfurt School were right to make the sociological claim that practical life under capitalism is governed by instrumental reason. I am not doing sociology but just considering whether we could get to certain conclusions by non-instrumental reasoning alone. (I do, however, suspect that we need to replace our currently existing reasoning with more specific forms of non-instrumental reasoning to avoid alienation.)

Although the Necessary Glue Claim is weaker than SEC, some of the points already made provide a sufficient case against this claim. In the previous section, I drew attention to the following styles of non-instrumental reasoning:

> *Constituents and Preconditions Reasoning*: One reasons from (1) an intrinsic[10] desire to do some complex activity A and (2) a belief about A's constitution and the preconditions for A-ing to (3) non-fundamental but intrinsic desires for A's constituents and (4) non-fundamental and extrinsic desires to establish the preconditions.
>
> *Specificationist Reasoning*: One reasons from (1) an intrinsic desire to do a generic activity-type and (2) a belief about what an especially desirable token of that activity would look like to (3) an intrinsic desire to do that token.
>
> *Tool-Aided Reasoning*: One reasons from (1) an intrinsic desire or intention to X, (2) the belief that tool T would be helpful for X-ing, to (3) a derivative but intrinsic desire to X via T.

[10] By 'intrinsic desire to X', I mean *desire to do X for its own sake*. Combined with my case for separating the fundamental/non-fundamental distinction from the non-instrumental/instrumental distinction in Sylvan (2018), the discussion here brings out that we really need a *threefold* distinction in valuing and value: (i) the for-its-own-sake/for-the-sake-of-something-else distinction, (ii) the fundamental/non-fundamental distinction and (iii) the non-instrumental/instrumental distinction. Threefold distinctions have been defended by others (see Tannenbaum (2010)), but the one here differs from anything in the existing literature.

> *'Sake'-Based Reasoning*: One reasons from (1) an intrinsic desire for X or intention to Y and (2) the belief that Z-ing is suitably related to X or Y-ing to (3) a non-fundamental but intrinsic desire to Z for X's sake or to Z for Y-ing's sake.

The first and third styles are structurally closest to alleged instrumental reasoning. It would be easiest to undermine the letter of the Necessary Glue Claim by appealing to them together with the view, already defended, that these forms of reasoning are not worth calling instrumental.

The third style easily replaces alleged instrumental reasoning. It is close enough, however, that one might not see it as grounding a sufficient case against the *spirit* of the Necessary Glue Claim. Properly appreciated, the first style is a more promising replacement. The attitude it takes toward smaller actions that are parts of some larger activity is genuinely different from the attitude that instrumental reasoning takes. It makes good sense to regard a person's desire for the constituents of an intrinsically worthwhile activity to be intrinsic and to think of the person as enjoying the parts in the same way they enjoy the whole. Indeed, the whole in this view is enjoyed *through* enjoying the parts. Contrast the instrumental attitude. Eating each bite as a means to bringing it about that one eats the sandwich is very different from eating each bite as a part of eating the whole sandwich. The first is not enjoyable, barring bizarre values. The second is enjoyable if the sandwich is good.

We can, I think, easily enough imagine replacing instrumental reasoning with either the first or third styles, though only the first embodies a different attitude toward life. A transition to either of these styles would be less radical than a transition to the second and fourth styles, however. The transition would be like the transition from greyscale to color, with the image otherwise remaining the same, and not like the transition from photography to painting or music. By contrast, shifting to the second or fourth styles would be more like changing the medium of practical thought.

Partly for this reason, it is harder to see how these transitions would work. The transition to the second cannot be done without a loss of important granularity, as far as I can see. While there are some brilliant re-imaginings of practical reasoning in Kolnai (1962), following him in regarding specification as the sole fundamental form of practical reasoning replaces the bones of action with meat. Practical reasoning needs a skeleton to move – this is Vogler's (2002) insight. Desiring to experience the *denouement* is not a more specific way of desiring to enjoy the story. So, we would need to combine this second style with the first. But the first alone captures what we might want from the second. For parts are not the only things we can regard as constituents of practical activities: specifications could be regarded as constituting tokens of activity types.

Replacing instrumental reasoning with 'sake'-based reasoning may seem an easier task, involving a single act of find-and-replace. But it will involve a reformatting of practical reasoning and probably a shift in values. Note that the only simple way to

replace an instrumental intention with a 'sake'-based intention will be to represent one as acting *for the sake of bringing about the activity toward which the instrumental intention is directed*. Acting for the sake of production will often seem perverse. Hence, these intentions may need to be dropped once their meaning is laid bare by the transition to 'sake'-based reasoning. But I think we must trust that an adequate alternative can be found if we want to represent an agent's activity as meaningful from her point of view (which I think is required for her activity to be fully rational).

In the next and final major section, I will say more about how I think this replacement should go. For now, we can rest assured that there are several ways to replace instrumental reasoning with non-instrumental reasoning which will not involve excessive loss of practical structure. Hence, we can reject the Necessary Glue Claim: there are other ways to hold practical reasoning together.

4 The non-instrumental structure of practical reason

Suppose one agrees that we should do without instrumental rationality. How then should we understand the non-instrumental structure of practical rationality? Our answer should be guided by some constraints from the previous sections:

> *Non-Alienation*: Transparent manifestations of practical rationality should not be alienating as such.
>
> *Necessary Intrinsic (but Derivative) Value*: (1) Necessarily, each manifestation of practical rationality should confer something of intrinsic value on the attitudes or actions that manifest it. (2) But this intrinsic value should not be *fundamental*, since rationality is not of *fundamental* value: it should be derived from a relation to a more fundamental value – just a *non-instrumental* relation.
>
> *Sufficient Generality*: The principles or pressures that underwrite the rationality of transitions should have sufficient generality: they shouldn't be so specific as to obscure rational similarities between different styles of transitioning.
>
> *Sufficient Granularity*: The principles or pressures that underwrite the rationality of transitions should not be so general as to render all fine structure of these transitions epiphenomenal.

Existing views which dispense with principles of instrumental rationality at the fundamental level violate some of these constraints.

On the one hand, views like Raz's (2005), Lord's (2018) and Kiesewetter's (2017) that seek to explain rationality by appealing to reasons threaten to disrespect the

fourth constraint. While these figures have error theories to explain away the apparent significance of structural relations, it would be nice to avoid giving an error theory. Parallel points apply to other views which seek to ground rationality in substantive normativity, like Anderson's (1993) fitting-attitudes account of rationality. The point would also apply to views which seek to privilege some specific forms of structural rationality, such as Kolodny's (2005) 'transparency account', which makes the Enkratic Principle the supreme principle of structural rationality.

On the other hand, the more structural alternatives in the literature violate other constraints. As I have already discussed, the specificationism of Kolnai (1962), Millgram (2001), Richardson (1994) and Wiggins (1975) violates the fourth constraint, depriving practical reasoning of its skeleton. As I have not already discussed, I think the constituents-and-preconditions view (perhaps this is Thompson's 2008 view) doesn't make clear sense of the first or second constraints. Here I would want to invert a thought from Wallace (2001) about cases in which one skillfully executes activities one doesn't value. Wallace thinks a theory of rationality should explain the *cleverness* displayed in these cases. But if rationality has the value I assume it has, I don't see why we should regard alienated cleverness as a manifestation of practical rationality. If the activity is meaningless from the agent's perspective and doesn't reflect her values, I see no rationality in intending the constituents of the activity. There may be great practical *skill* on display, but that is not our topic here. Notice finally that the complaint isn't a substantive complaint: there might be conclusive substantive objections to an agent's values, but an act might still be rational in the relevant sense relative to those values.

Although I don't think the constituents-and-preconditions approach works on its own, I suspect that this style of reasoning will remain part of the best alternative. As I have hinted, my preferred approach replaces instrumental relations with 'for the sake of' relations. What is at the end of the chain of such relations is some ultimate value one holds dear, where examples of proper values might include equality, liberty, truth, happiness, wisdom.

One's ultimate values directly rationalize one's final intentions. These intentions are formed in light of one's conception of those ultimate values. Valuing equality and liberty together might, for example, rationally require having an intention to relate to others as democratic equals, where some background conception of equality and liberty leads one to treat democracy as the social arrangement that best embodies these values. While these intentions are final, they are adopted *for the sake of* the ultimate values. One might finally intend to relate to others democratically, for example, *because* it respects equality and liberty.

Practical reasoning doesn't end with these intentions: this is the structural insight of instrumentalism. But the relationship between these intentions and derived intentions to do specific actions won't be instrumental. Here constituents-and-preconditions reasoning returns in a subordinate role. The fundamental rationalizing relations will remain 'for-the-sake-of' relations, but these relations will hold between one's ultimate

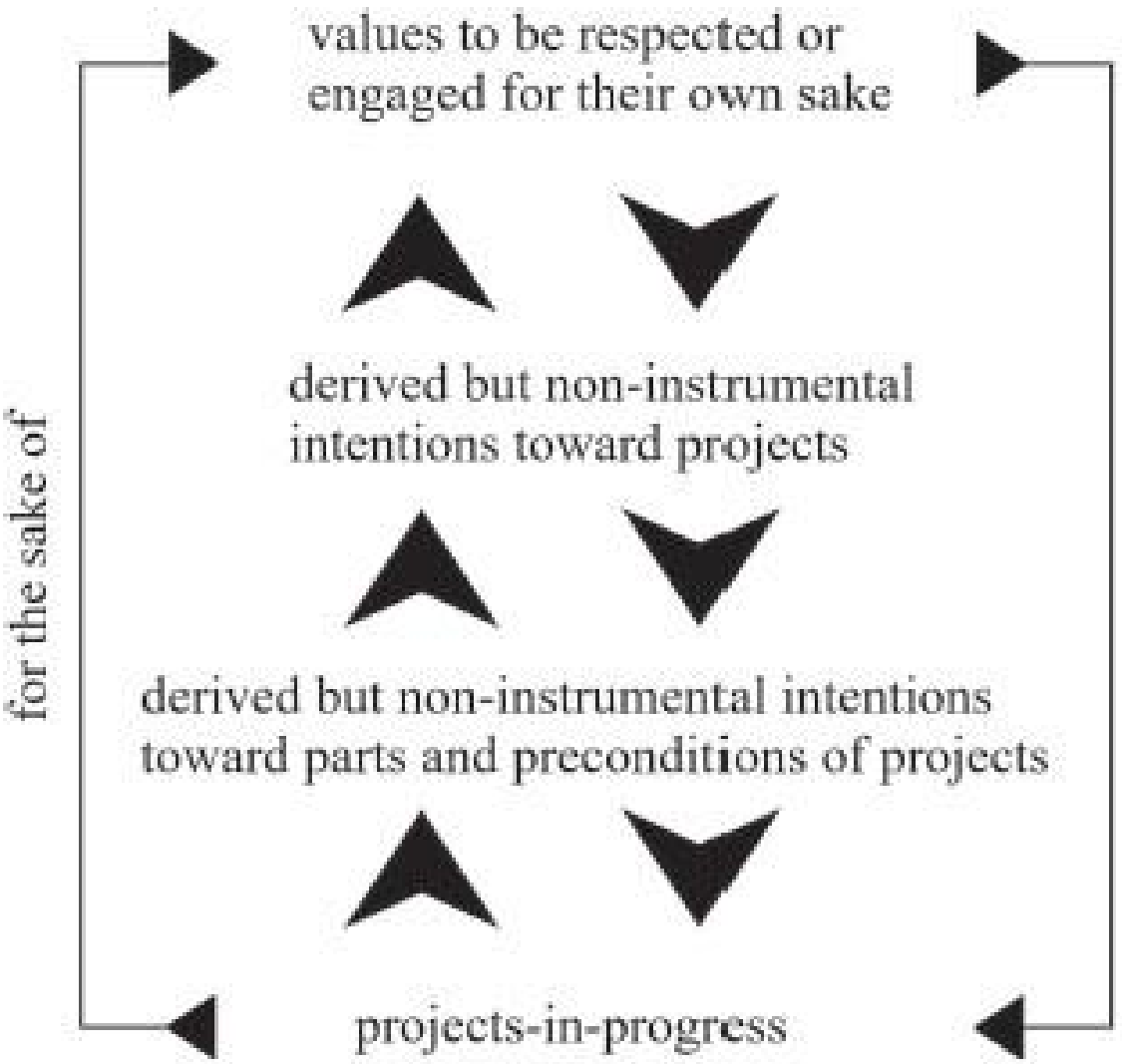

values to be respected or
engaged for their own sake
derived but non-instrumental
intentions toward projects
derived but non-instrumental intentions
toward parts and preconditions of projects
projects-in-progress
for the sake of
rationalize

Figure 32.1

and narrower intentions *through* the structure of parts and preconditions. In a diagram, the view suggests that a rational agent's practical mind will be structured as in Figure 32.1.[11]

With this picture in mind, let's walk through our checklist of constraints.

First, this picture satisfies the Non-Alienation constraint *unless* one thinks that avoiding alienation requires adopting specific values; I will return to this question momentarily. What seems clear is that having one's practical thought held together by this kind of structure is sufficient to give it meaning by one's own lights.

Second, I take it that the foregoing fact gives as much value to structural rationality as one could reasonably expect. For-the-sake-of relations can, I suggest, transmit the meaningfulness of highest-level values to the lower levels of practical thought (e.g., intentions directed at parts and preconditions). Such meaningfulness has some measure of ultimate value, and the lower levels share in this value by bearing for-the-sake-of relations to the top level.

Finally, the picture has a 'goldilocks' degree of generality which satisfies the third and fourth constraints. Although there might be special for-the-sake-of relations needed to understand the structure of specific values, a general account of rationality can omit such details. We don't, however, want to omit the steps that take many rational agents from the top to the bottom level. Perhaps we can imagine agents who move directly from the top level to the bottom level through one exercise of *phronesis*. But for many agents, practical rationality is often harder work, essentially including extra steps. Some or even most of the rationality of such agents' transitioning would drain out if these steps were skipped.

The one question that remains is a relative of Parfit's (1984) question about whether some desires are intrinsically irrational in Scanlon's (1998) narrow sense. One might wonder whether some values would necessarily have deficient meaning in a correspondingly narrow sense (i.e., meaningfulness from the subject's perspective). Smith (1995) and Markovits (2014) believe that structural rationality excludes immorality. I'm not sure. But I leave open whether some substantive values are structurally excluded. Indeed, I am tempted to opt for constitutivism and hold that being a being of a certain kind necessarily involves having values of a certain cast. There are values for the ultimate sake of which we cannot meaningfully act. If Arendt was right, *usefulness* is an example.

My suspicion is that moral requirements will not follow from the structural constraints on for-the-sake-of relations. This is my spinoff on Velleman's (1992) thought that full-blooded agents aren't necessarily 'squares'. I would deny that morality is *at odds* with practical reason. But I don't yet see that morality in the narrow sense falls

[11] The picture doesn't depict one important ingredient – that is, non-normative beliefs about the relations that must be in place for an attitude or act to be carried out for the sake of a value. These beliefs will replace the beliefs about instrumental relations that figure in the 'standard story of action'.

I should also note that I leave open whether the values and intentions in the picture are cognitive or conative states. I hope the picture is consistent with a cognitivist view about practical reasoning

out of practical reason. A wider Aristotelian story might be more plausible, or there might be better constitutivist pictures, such as Katsafanas's (2013) Nietzschean picture or the neo-Aristotelian picture that Wood (1999) and Hurka (1993) ascribe to Marx. All of these *and* the Kantian pictures are consistent with my overall view, barring further arguments. The disagreements can be seen as disagreements about what can coherently terminate a 'for-the-sake-of' chain. It is beyond the scope of this chapter to resolve these disagreements.

5 The non-instrumental unity of reason

I conclude by drawing attention to a final virtue of the view, which further distinguishes it from views featuring some fundamental instrumental requirement: it allows us to see reason as unified across its practical and epistemic manifestations. As I argued in Sylvan (2014, 2018, 2020) and others have argued,[12] instrumentalist views do not provide a tolerable unification of epistemic and practical normativity. Hence, if one rejects instrumentalism but still accepts a fundamental requirement of instrumental rationality, the outcome will be disunity: since epistemic rationality is never instrumental rationality, it will be fundamentally unlike much of practical rationality on standard views. Thankfully, there is a non-instrumentalist unification that is as comprehensive as the attempted instrumentalist unifications of pragmatists like James (1896/1979) and Rinard (2015, 2017), instrumentalists like Foley (1987, 1992) and consequentialists like Pettigrew (2016).

But before turning to this alternative unification, I first want to emphasize a different kind of unity that has already emerged but not received explicit comment. As I said earlier, a key insight of Vogler (2002) is that most practical reasoning is not reasoning about ultimate values (*pace* the specificationists) but rather a sequence proceeding from big-picture values to small-scale intentions and actions. What norm gives order to this sequence? For Vogler, it is the instrumental principle. But unless one is either an instrumentalist or a proponent of the view that all ultimate value is 'to be promoted', this norm will seem fundamentally unlike the norm that governs embrace of ultimate values. Hence, one will get separate hypothetical and categorical imperatives, making practical reason fundamentally divided.

I upheld Vogler's basic insight but claimed that the norm that governs the sequence from ultimate values to small-scale intentions and actions is of the same kind as the norm that governs embrace of ultimate values. Ultimate values are embraced with fitting attitudes. What makes it fitting to embrace a value is the same as what makes

as well as a divided view that features both cognitive and irreducibly conative states. But at least if 'Humean' implies 'instrumentalist', the divided view will be divorced from the Humean theory of motivation.

[12] See Fumerton (2001), Kelly (2003) and Berker (2013). See Foley (1987, 1992) for the clearest example of an attempt at instrumental unity and the primary target of Fumerton and Kelly.

it fitting to manifest one's embrace in practice, by acting for the sake of the value. Hence the structure of practical reason is the structure of embracing, thinking, and acting for the sake of value (which is carried out in the messy empirical realm and is hence a mess of complicated steps). In this view, both structural rationality in the narrow sense (i.e., rationality *relative* to other attitudes) and the rationality attaching to one's values have the same normative ground.

Although practical reason is in this way unified, it is worth distinguishing two kinds of values that could structure one's practical mind. On the one hand, there are values to be *respected*, where respect is understood as imposing a *deliberative constraint* not to disrespect the value by acting in certain ways (e.g., in the case of personhood, by violating rights). But not all value is primarily to be respected in this negative sense. Much value is to be engaged with, where engagement is positive.[13] By creating or taking pleasure in art, for example, we engage with aesthetic value. Some value to be respected also merits engagement: persons are not just to be respected but also loved, as Kant emphasized in *The Metaphysics of Morals*. Engagement and its constitutive norms are the answer to alienation and the instrumental attitude to value embodied in a production-first view. For respect alone seems an inadequate source of meaning (unless one is as stuffy as legend portrays Kant).

With this distinction in mind, we can more easily explore how the view developed previously combines with the view of epistemic rationality from Sylvan (2014, 2018, 2020) to yield a unified picture. In those earlier works, I suggested that rational belief derives epistemic value from manifesting *respect for truth*. This view generates a picture (see _Figure 32.2]]) like the previous one, though it inserts a specific value (truth) in place of the placeholder.[14]

So far, this view is stereotypically Kantian, deriving everything from respect. But Kant reserved a place for positive forms of valuing (e.g., love). Where in the theoretical domain might one expect to find these? Not at the core of epistemic rationality in the narrowest sense, I think. Narrow epistemic rationality – that is, the constitutive rationality of belief and other stative theoretical attitudes – is fundamentally negative, though iffy positive requirements might arise if one seeks to settle some particular questions.[15]

Yet besides occupying states like belief, we engage in activities like inquiry and theoretical reasoning. Once we appreciate this point, we may find a place for something

[13] I take the term 'engagement' from Raz (2002), who defends a similar view about value.

[14] The value at the top level is one to which I think we are essentially committed in virtue of being cognizers; hence 'value' is also used in an internal sense here, to mean a value *of the cognizer*. I am unsure whether we believe *for the sake of* truth (see Sosa 2000, 2015: Ch.2 for worries). But we do constrain our doxastic attitudes for the sake of accuracy. Hence a doxastic attitude could manifest *something* which is for the sake of truth (respect). Here it might be better to say that our theoretical reason manifests respect for truth through belief (adopted non-voluntarily). These points suggest that some adjustment to the interpretation of the arrows on the left might be needed.

[15] See Nelson (2010) for the first point and Sylvan (2016) for the qualifier.

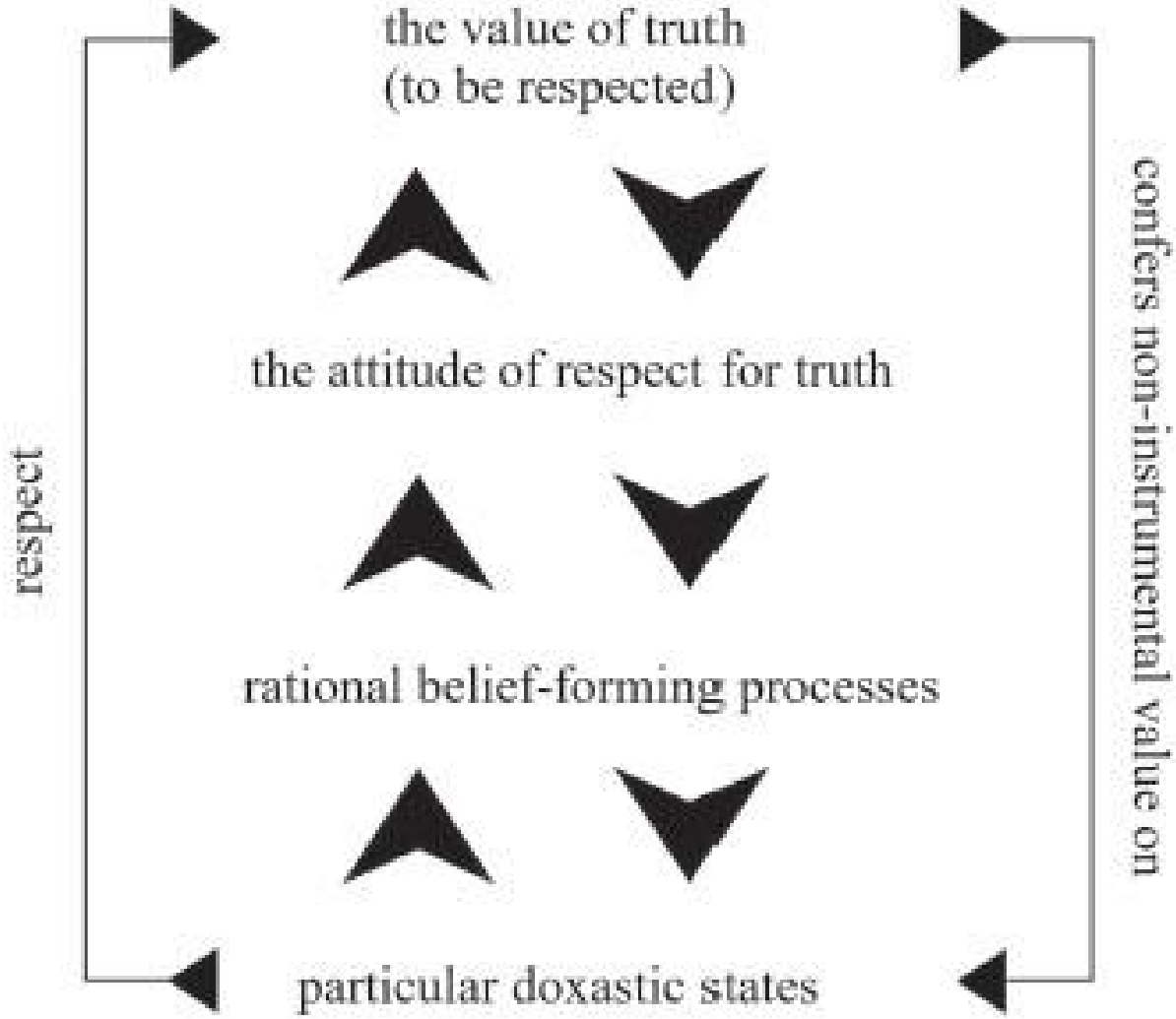

the value of truth
(to be respected)
the attitude of respect for truth
rational belief-forming processes
particular doxastic states
confers non-instrumental value on
respect

Figure 32.2

more positive, which would yield fuller symmetry between the epistemic and the practical. Plausibly, inquiry is an attempt to engage with reality, by opening one's mind to the facts and seeking to perceive them aright. If there is reason of the right kind to engage with truth for its own sake – a reason which would give force to the criticism of *uninquisitiveness* – then truth will merit engagement as well as respect.

One might wonder whether commands of engagement and respect might conflict or represent norms of different kinds, as Friedman (forthcoming) suggests in discussing the epistemic and the 'zetetic'. But I suspect we can restore harmony if we model the zetetic *not* on instrumental practical rationality (as Friedman assumes) but on non-instrumental structural rationality. Friedman assumes that because inquiry is an activity, it generates instrumental pressures. One can block the argument by purging the practical of anything fundamentally instrumental.

It is beyond the scope of this chapter to show that the zetetic has the same structure as practical rationality as I've understood it. But the challenge to the unity of the theoretical exists only if the zetetic is governed by instrumental rationality. I agree with Friedman that the study of theoretical reason goes beyond the study of the narrowly epistemic (i.e., the constitutive norms of stative theoretical attitudes). Given this point, the picture sketched previously together with the picture I've already defended in epistemology leads to a unification of rationality.[16]

[16] For helpful feedback on earlier versions of this chapter, I thank Jonathan Dancy, Alex Gregory, Brian McElwee, Sarah Paul, Christian Piller, John Schwenkler, David Sosa, Jonathan Way, Daniel Whiting, Fiona Woollard and other members of audiences at the University of Southampton and the University of Texas at Austin.

Bibliography

Adorno, T. and Horkheimer, M. 1944/1979. *The Dialectic of Enlightenment*. London: Verso Books.

Anderson, E. 1993. *Value in Ethics and Economics*. Cambridge, MA: Harvard University Press.

Anscombe, G. E. M. 1957. *Intention*. Oxford: Blackwell.

Arendt, H. 1958. *The Human Condition*. Chicago: University of Chicago Press.

Audi, R. 2001. *The Architecture of Reason*. Oxford: Oxford University Press.

Bengson, J. and Moffett, M. 2012. 'Non-Propositional Intellectualism' in J. Bengson and M. Moffett (eds.) *Knowing How: Essays on Knowledge, Mind, and Action*. Oxford: Oxford University Press.

Benhabib, S. 1994. 'The Critique of Instrumental Reason' in S. Žižek (ed.) *Mapping Ideology*. London: Verso Books.

Berker, S. 2013. 'Epistemic Teleology and the Separateness of Propositions' *Philosophical Review* 122: 337–393.

Bratman, M. 1987. *Intention, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press.

Broome, J. 1999. 'Normative Requirements' *Ratio* 12: 398–419.

Broome, J. 2005. 'Does Rationality Give Us Reasons?' *Philosophical Issues* 15: 321–337.

Broome, J. 2007. 'Does Rationality Consist in Correctly Responding to Reasons?' *Journal of Moral Philosophy* 4: 349–374.

Broome, J. 2013. *Rationality Through Reasoning*. Malden, MA: Blackwell.

Brunero, J. 2012. 'Instrumental Rationality, Symmetry, and Scope' *Philosophical Studies* 157: 125–140.

Brunero, J. 2020. *Instrumental Rationality*. Oxford: Oxford University Press.

Carter, J. A. and Jarvis, B. 2012. 'Against Swamping' *Analysis* 72: 690–699.

Dancy, J. 2018. *Practical Shape*. Oxford: Oxford University Press.

Dreier, J. 1993. 'Structures of Normative Theories' *The Monist* 76: 22–40.

Dreier, J. 1996. 'Rational Preference: Decision Theory as a Theory of Practical Rationality' *Theory and Decision* 40: 249–276.

Fogal, D. Forthcoming. 'Rational Requirements and the Primacy of Pressure.' *Mind*.

Foley, R. 1987. *The Theory of Epistemic Rationality*. Cambridge, MA: Harvard University Pres.

Foley, R. 1992. *Working Without a Net*. Oxford: Oxford University Press.
Friedman, J. Forthcoming. 'The Epistemic and the Zetetic.' *Philosophical Review*.
Fumerton, R. 2001. 'Epistemic Justification and Normativity' in M. Steup (ed.) *Knowledge, Truth, and Duty*. Oxford: Oxford University Press.
Gauthier, D. 1987. *Morals by Agreement*. Oxford: Clarendon Press.
Gorz, A. 1989. *The Critique of Economic Reason*. London: Verso Books.
Hampton, J. 1998. *The Authority of Reason*. Cambridge: Cambridge University Press.
Harman, G. 1976. 'Practical Reasoning' *Review of Metaphysics* 29: 431–463.
Heidegger, M. 1954/1977. *The Question Concerning Technology and Other Essays*. New York: Harper and Row.
Hill, T. 1973. 'The Hypothetical Imperative' *Philosophical Review* 82: 429–450.
Horkheimer, M. 1947. *The Eclipse of Reason*. Oxford: Oxford University Press.
Horkheimer, M. 2012. *The Critique of Instrumental Reason*. London: Verso Books.
Hurka, T. 1993. *Perfectionism*. Oxford: Oxford University Press.
Hurka, T. 2001. *Virtue, Vice and Value*. Oxford: Oxford University Press.
Hurley, P. 2018. 'Consequentialism and the Standard Story of Action.' *Journal of Ethics* 22: 25–44.
James, W. 1876/1979. 'The Will to Believe' reprinted in *The Will to Believe and Other Essays*. Cambridge, MA: Harvard University Press.
Johnson, M. 2005. *Aristotle on Teleology*. Oxford: Clarendon Press.
Jones, W. 1997. 'Why Do We Value Knowledge?' *American Philosophical Quarterly* 34: 423–439.
Kant, I. 1785/2012. *Groundwork for the Metaphysics of Morals* in *Practical Philosophy*, eds. M. Gregor and J. Timmerman. Cambridge: Cambridge University Press.
Katsafanas, P. 2013. Agency and the Foundations of Ethics: Nietzschean Constitutivism. *Oxford: Oxford University Press.*
Kelly, T. 2003. 'Epistemic Rationality as Instrumental Rationality: A Critique' *Philosophy and Phenomenological Research* 66: 612–640.
Kiesewetter, B. 2017. *The Normativity of Rationality*. Oxford: Oxford University Press.
Kolnai, A. 1962. 'Deliberation Is of Ends' *Proceedings of the Aristotelian Society* 62: 195–218.
Kolodny, N. 2005. 'Why Be Rational?' *Mind* 114: 509–563.
Kolodny, N. 2007. 'How Does Coherence Matter?' *Proceedings of the Aristotelian Society* 107: 229–263.
Kolodny, N. and Brunero, J. 2018. 'Instrumental Rationality' in *The Stanford Encyclopedia of Philosophy*. Palo Alto: Stanford University Press.
Korsgaard, C. 1983. 'Two Distinctions in Goodness' *Philosophical Review* 92: 169–195.
Korsgaard, C. 1997. 'The Normativity of Instrumental Reason' in G. Cullity and B. Gaut (eds.) *Ethics and Practical Reason*. Oxford: Clarendon Press.
Korsgaard, C. 2009. *Self-Constitution*. Oxford: Oxford University Press.
Lavin, D. 2013. 'Must There Be Basic Action?' *Nous* 47: 273–301.
Lord, E. 2018. *The Importance of Being Rational*. Oxford: Oxford University Press.

Marcuse, H. 1964. *One-Dimensional Man*. London: Routledge.
Markovits, J. 2014. *Moral Reason*. Oxford: Oxford University Press.
Marx, K. 1844/1988. *The Economic and Philosophical Manuscripts of 1844*. Buffalo, NY: Prometheus Books.
Millgram, E. 1997. *Practical Induction*. Cambridge, MA: Harvard University Press.
Millgram, E. (ed.) 2001. *Varieties of Practical Reasoning*. Cambridge, MA: The MIT Press.
Nelson, M. T. 2010. 'We Have No Positive Epistemic Duties' *Mind* 119: 83–102.
Nozick, R. 1993. *The Nature of Rationality*. Princeton: Princeton University Press.
Parfit, D. 1984. *Reasons and Persons*. Oxford: Oxford University Press.
Pettigrew, R. 2016. *Accuracy and the Laws of Credence*. Oxford: Oxford University Press.
Pettit, P. 1989. "Consequentialism and Respect for Persons" *Ethics* 100: 116–126.
Portmore, D. 2007. 'Consequentializing Moral Theories' *Pacific Philosophical Quarterly* 88: 39–73.
Portmore, D. 2011. *Commonsense Consequentialism*. Oxford: Oxford University Press.
Prichard, H. A. 1937. 'Moral Obligation' Reprinted in McAdam, J. 2002 (ed.) *H.A. Prichard: Moral Writings*. Oxford: Oxford University Press.
Railton, P. 1984. 'Alienation, Consequentialism, and the Demands of Morality' *Philosophy and Public Affairs* 13: 134–171.
Raz, J. 2002. *Engaging Reason*. Oxford: Oxford University Press.
Raz, J. 2005. 'The Myth of Instrumental Rationality' *Journal of Ethics and Social Philosophy* 1: 1–28.
Raz, J. 2011. *From Normativity to Responsibility*. Oxford: Oxford University Press.
Raz, J. 2016. 'Value and the Weight of Reasons' in B. Maguire and E. Lord (eds.) *Weighing Reasons*. Oxford: Oxford University Press.
Richardson, H. 1994. *Practical Reasoning About Final Ends*. Cambridge: Cambridge University Press.
Ridley, A. 2018. The Deed Is Everything: Nietzsche on Will and Action. *Oxford: Oxford University Press.*
Rinard, S. 2015. 'Against the New Evidentialists' *Philosophical Issues* 25: 208–223.
Rinard, S. 2017. 'No Exception for Belief' *Philosophy and Phenomenological Research* 94: 121–143.
Scanlon, T. M. 1998. *What We Owe to Each Other*. Cambridge MA: Harvard University Press.
Scanlon, T. M. 2007. 'Structural Irrationality' in G. Brennan, R. Goodin, F. Jackson and M. Smith (eds.) *Common Minds: Themes from the Philosophy of Philip Pettit*. Oxford: Clarendon Press.
Schroeder, M. 2004. 'The Scope of Instrumental Reason' *Philosophical Perspectives* 18: 337–364.
Schroeder, M. 2007a. *Slaves of the Passions*. Oxford: Oxford University Press.

Schroeder, M. 2007b. 'Teleology, Agent-Relative Value, and "Good"' *Ethics* 117: 265–295.
Schwenkler, J. 2019. *Anscombe's Intention: A Guide*. Oxford: Oxford University Press.
Setiya, K. 2007. 'Cognitivism about Instrumental Reason' *Ethics* 117: 649–673.
Smith, M. 1995. 'Internal Reasons' *Philosophy and Phenomenological Research* 55: 109–131.
Sosa, E. 2000. 'For the Love of Truth?' in A. Fairweather and L. Zagzebski (eds.) *Virtue Epistemology*. Oxford: Oxford University Press.
Sosa, E. 2015. *Judgment and Agency*. Oxford: Oxford University Press.
Stanley, J. and Williamson, T. 2001. 'Knowing How' *Journal of Philosophy* 98: 411–444.
Sylvan, K. 2014. *On the Normativity of Epistemic Rationality*. PhD Thesis, Rutgers University, New Brunswick.
Sylvan, K. 2016. 'The Illusion of Discretion' *Synthese* 193: 1635–1665.
Sylvan, K. 2018. 'Veritism Unswamped' *Mind* 127: 381–435.
Sylvan, K. 2020. 'An Epistemic Non-Consequentialism' *Philosophical Review* 129: 1–51.
Tannenbaum, J. 2010. 'Categorizing Goods' in R. Shafer-Landau (ed.) *Oxford Studies in Metaethics 5*. Oxford: Oxford University Press.
Thompson, M. 2008. *Life and Action*. Cambridge, MA: Harvard University Press.
Velleman, D. 1992. 'The Guise of the Good.' *Nous* 26: 3–26.
Vogler, C. 2002. *Reasonably Vicious*. Cambridge, MA: Harvard University Press.
Wallace, R. J. 2001. 'Normativity, Commitment, and Instrumental Reason' *Philosophers' Imprint* 1(3): 1–26.
Way, J. 2012. 'Explaining the Instrumental Principle' *Australasian Journal of Philosophy* 90: 487–506.
Weber, M. 1921/1968. *Economy and Society*. New York: Bedminster Press.
Wiggins, D. 1975. 'Deliberation and Practical Reason' *Proceedings of the Aristotelian Society* 76: 29–51.
Williams, B. 1981. 'Internal and External Reasons' in *Moral Luck*. Cambridge: Cambridge University Press.
Wood, A. 1999. *Karl Marx*. London: Routledge.
Zagzebski, L. 2000. 'From Reliabilism to Virtue Epistemology' *Proceedings of the 20th World Congress of Philosophy* 5: 173–179.

33. Rationality, regret, and choice over time

Introduction

As less than perfect temporally extended agents, most of us are familiar with the experience of making a choice or series of choices that we later look back on with regret. This article will look at three sorts of scenarios that illuminate this phenomenon and review some philosophical arguments concerning whether and, if so, how agents in the described predicaments are going wrong. The three sorts of scenarios can be roughly described as follows: (1) cases in which the experience of temptation prompts a temporary preference reversal, (2) cases in which the vagueness of a goal prompts procrastination, and (3) cases in which shuffling between incommensurable alternatives results in a gratuitous cost.[1] Each sort of scenario is associated with a preference structure that raises a challenge with respect to effective choice over time, as well as associated worries regarding the agent's rationality. Relatedly, the scenarios seem to invite the sort of regret that goes along with a sense that one somehow failed oneself, which is the sort of regret on which I will here be focusing.

[1] These sorts of scenarios, their relationship to one another, and the debates surrounding them, are extensively discussed in my prior work in this area. See especially (Andreou 2012), (Andreou 2014), (Andreou 2015), and (Andreou 2016). My discussion here distills, with some revisions and recasting, some of the highlights of these discussions.

Going wrong

In discussing whether and how things are going wrong in the sorts of scenarios mentioned above, my focus will be on failures of *instrumental rationality*, where instrumental rationality is to be understood as providing *hypothetical imperatives* (and only hypothetical imperatives), and hypothetical imperatives are to be understood as imperatives of reason that are relative to the agent's ends or concerns. Candidate imperatives of the relevant sort include, for example, "If you want stay on his good side, compliment him daily," and "If you want a comfortable retirement, don't keep putting off saving." I will not delve into the question of whether, when it comes to guidance regarding action, rationality provides anything other than hypothetical imperatives.[1] There seems to be room, in the scenarios of interest, for an agent to fail relative to his own ends and concerns, and it is the possibility of such failure that will here concern me.

Notice that, as instrumental rationality is here understood, a failure of instrumental rationality can be grounded in a false belief and so ultimately be due to a theoretical error rather than a (distinctively) practical error. Consider, to pick up on an (ever-so-slightly adjusted) example from Bernard Williams's work (1979), the case of an agent who inadvisably drinks the contents of the glass in front of him thinking it is gin and tonic when it is actually petrol and tonic. This qualifies as a failure of instrumental rationality in the relevant sense. I should emphasize that I do not accept David Hume's view (1978 [1739–40]) that all failures of rationality are grounded in some sort of misrepresentation. To the contrary, I think that some failures of rationality and, in particular, some failures of instrumental rationality, reflect other shortfalls, such as, for example, poor managerial skills. This particular shortfall is, in my view, the problem in many cases involving patterns of choice that are problematic relative to the agent's ends or concerns, including, for instance, the costly cases of shuffling between incommensurable alternatives that will be discussed later in this article.

[1] For some contemporary classics defending the view that rationality goes beyond providing just hypothetical imperatives, see, for example, (Korsgaard 1996), (Foot 2001), and (Smith 1994).

Temptation and temporary preference reversals

Regret is often associated with and illustrated via cases involving temporary preference reversals, such as the following. Suppose I've been invited to a party that promises to be intoxicatingly fun. The intoxicating fun will be achieved, at least in part, by my having at least three shots of liquor. Knowing that having more than three shots will result in my having a painful hangover the next day, I plan on having three shots and no more, but I am somewhat anxious, since I know from experience that, when I am at an intoxicatingly fun party, I tend to think that the added fun of a fourth shot is well worth the terrible hangover that I'll have to deal with the next day. Not surprisingly, once I am at the party, I abandon my plan, have a fourth shot, and suffer the next day; I later see my behavior at the party as self-defeating and regret having had a fourth shot, even after the pain of my hangover has subsided.

This case raises some interesting issues. Notice first that it might be wondered whether what we have here is just a conflict between different 'time slice selves' with different values and so different ends. If this is correct, then there is, it seems, no reason for the 'middle slice,' who favors more fun now even at the expense of pain suffered by a later self tomorrow, to deviate from her values in order to conform to her earlier or later time slice's values; at least this seems true from the point of instrumental rationality. But then it seems like there is no failure of instrumental rationality on the part of the agent in this case; indeed, it seems like there is no sufficiently unified temporally extended agent for instrumental rationality to advise. Though instrumental rationality can provide advice appropriate to each time slice's values, it cannot arbitrate between different value systems.

This way of understanding the case does not take the later self's regret at face value. The later self seems to be assuming that there is an enduring I in the scene and that the preference reversal in the face of temptation is to be understood as a lapse that goes against the enduring I's values. There are a variety of interesting philosophical positions that are relevant to this alternative way of understanding cases involving preference reversals in the face of temptation.

Consider first the idea that temporary preference reversals are due to *hyperbolic discounting*. When the evaluation of future options is shaped by hyperbolic discounting, the value associated with appealing options is discounted (i) at a rate that depends on how long one must wait before the appealing option becomes available, and (ii) in a way that makes the attractiveness of appealing options spike sharply when they

become imminent. Hyperbolic discounting can thus result in a last-minute switch in favor of an instance of indulgence even if the agent's general stance on the value of fun versus pain avoidance remains constant.[1] Let's return to the intoxicatingly fun party. It might be that, even when I want to go on to my fourth shot, it is not because I have now become convinced that I should turn over a new leaf and front-load fun even at the expense of later, greater pain. I might still think that, generally speaking, I should exercise a habit of showing restraint. It's just that, right now, with the attractiveness of the fun on offer looming so large because of its imminence, I currently favor making an exception at this point and showing restraint in the future. If this is the case, and my preference reversal is due not so much to my values changing as to the pull of an option increasing due to the "advantage of its situation," then maybe we should allow that there is an enduring agent in the scene and that the preference reversal really is a lapse induced by the experience of temptation.[2]

According to a related line of thought, an agent that experiences a preference reversal in the face of temptation need not see herself as deeply fragmented. She can review the different preference structures in play over time and think about whether one figures as the best candidate to speak for her as a whole.[3] For example, if an agent that has just had a preference reversal anticipated the reversal and expects, based on past experience, that she will regret it if she acts on her current ranking, she may seriously consider the possibility that her current ranking is not as well suited to speak for her temporally extended self as her past and future ranking. This may, of course, lead her to revise her current ranking, and, indeed, it may be that she is rationally required to revise her current ranking if she determines that it does not speak for herself as a whole. If this process of review and revision can occur, then, even if it doesn't always occur, it seems like the agent's seeing herself as temporally extended is vindicated. There really can be something that counts as the stance that speaks for her as a whole, and instrumental rationality can be accountable to that stance.

According to a somewhat different line of thought, what needs to be identified in potential cases of temptation is not a stance that speaks for the agent as a whole but a stance that is not skewed by the experience of temptation.[4] The problem with seeking a stance that speaks for the agent as a whole is that preference reversals prompted by temptation need not be temporary. To avoid cognitive dissonance and bolster self-esteem, an agent who gives in to temptation may actually change his values so as to rationalize his lapse. Consider, for example, an agent who initially values good health, but who, faced with temptation and the threat of having to admit to failure, develops and maintains a newfound allegiance to "living fast and dying young" (assuming the latter is the price of the former). To avoid such rationalizations, a

[1] For an influential discussion on temporary preference reversals and hyperbolic discounting, see (Ainslie 2001).

[2] The quoted phrase is from (Hume 1978 [1739–40], 416).

[3] See (Bratman 2014) for a well-developed position along these lines.

[4] This idea, as well as those in the remainder of this paragraph are developed in (Holton 2009).

sensible agent may need to refuse to reconsider rankings he made 'in a cool hour,' and, barring unanticipated changes, resolutely stick to these rankings even if he is inclined to reconsider.

Interestingly, it might be wondered why the agent's value switch should be frowned upon just because it is motivated by the agent's desire to avoid compromising his self-esteem. If, in particular, instrumental rationality is neutral between a commitment to healthy living, on the one hand, and a commitment to living fast and (if necessary) dying young, on the other, why should one stick with the former value system when doing so is difficult and alternative value systems are, for all that has been said so far, rationally permissible? Relatedly, why describe the agent's newfound stance as skewed by the experience of temptation rather than as prompted by a more vivid sense of what the agent's initial values require him to give up and by the revelation that he is not ready to make the sacrifice? I don't know if this challenge can be adequately addressed. In my view, worries regarding the agent's instrumental rationality are more solidly grounded in cases where the agent looks back on her choice(s) with regret and the sort of enduring reevaluation I have been considering does not arise (Andreou 2014).

Vague goals and procrastination

I turn next to the second sort of scenario mentioned in the introduction and consider cases in which regret follows upon procrastination. In some such cases, procrastination may involve repeated irresoluteness due to temptation-induced preference reversals, but procrastination can also occur without any preference reversals (Andreou 2007a). Having reviewed some central ideas concerning temptation and preference reversals in the preceding section, I will focus, in this section, on cases of procrastination that do not involve any changes in one's rankings over time. Such cases are easily prompted by vague goals, which, as will become apparent, can generate stable but cyclic preferences that, if followed, result in regrettable delay.[1]

Consider the following case: Suppose that I have the goal of saving for a comfortable retirement. There may be no sharp boundary between the amount of funds that would support a comfortable retirement for me and the amount of funds that would fail to support a comfortable retirement for me. Relatedly, it may be that, given the frequency and amount of my pay, there is no pay period which is such that whether or not I save the available part of my pay during that particular period will make or break my prospect of a comfortable retirement. In such a case, it may be that no matter how much or how little I have saved by pay period x, I prospectively, and at period x, and retrospectively favor (re)initiating saving at period x+1 over (re)initiating saving at period x. And this may hold for every pay period. I may thus repeatedly delay saving and ultimately fail to achieve my goal. Since my preferences are stable over time, it is not the case that, looking back on some particular pay period, I will regret not having saved during that pay period. Before, during, and after that pay period, I will see that potential contribution as trivial. Still, I will presumably regret not having saved often enough to achieve my goal.[2]

Although my preferences in this case are stable over time, following them at each time period can lead to regret because they are cyclic. Though there is no disagreement between my 'time slices,' my preferences over certain options form a loop, not a linear ordering of the options (ranging from most preferred to least preferred, with, perhaps, some ties). If, leaving tangential complications aside, we use *e_ n* to stand for "make n exception(s) to adding the available part of my pay to my retirement account" and

[1] For a relevant discussion regarding procrastination and cyclic preferences, see (Andreou 2007a). For a relevant discussion regarding vague goals and rational choice, see (Tenenbaum and Raffman 2012) and (Tenenbaum in press).

[2] See (Andreou 2014) for an extensive discussion on regret in cases with this structure.

assume that there are 2000 pay periods, the loop in question can be roughly depicted as in _Figure 33.1]].

Things would not be too bad if all the possible outcomes were good, but this is not the case. While some of the possible outcomes are good, including, in particular, those in which I comfortably retire after steadily saving except for the occasional happy splurge, others, including those in which I must suffer an extremely uncomfortable retirement, are very bad (by my own lights).[3] Moreover, it is precisely such a bad outcome that I will be led to if I follow my preferences at each pay period and refrain from contributing the available part of my pay to my retirement account (with the recognition that, however many other exceptions I make in all, making an exception during this pay period will not make or break my prospect of a comfortable retirement).

According to a common view in the literature, cyclic preferences are rationally impermissible, and so an agent with cyclic preferences cannot escape criticism. The idea is that, even if there are no objectively required ends or rankings, an agent's preferences must still satisfy certain structural requirements, such as the requirement that they provide a linear ordering of the options.[4] According to a related but distinct view, rationality does not prohibit cyclic preferences but can only advise an agent with respect to non-cyclic preferences, and so an agent with cyclic preferences over the available options cannot seek guidance from rationality.[5]

There is also an alternative line of thought according to which cyclic preferences are sometimes permissible, but rationality requires an agent with cyclic preferences to proceed with caution. More specifically, the agent must think about whether following his cyclic preferences will threaten his 'big picture' goals and/or lead him to defeat his own purposes by, for example, leading him to an outcome that he deems unacceptable when one or more outcomes that he deems acceptable are available.[6] Where this threat exists, the agent must avoid going wrong by, for example, forming a good plan or policy and sticking to it or by relying on habits that prompt him to effectively use his discretion and show restraint in good time.

I favor this alternative line of thought for two reasons. First, unlike the stance that cyclic preferences are rationally impermissible, it allows for the idea that instrumental

[3] For a discussion of the significance, with respect to instrumental rationality, of cases involving preferences loops containing options in different leagues or categories, see (Andreou 2015).

[4] This view accords with interpreting the axiom of transitivity (which precludes, among other things, cyclic preferences) as a requirement of rationality rather than as just a precondition for the applicability of a certain way of representing preferences. The most prominent argument in favor of the view that rational preferences must be transitive (and so non-cyclic) is the "money-pump argument" (Davidson, McKinsey, and Suppes 1955). For an influential response, see (Schick 1986). I develop a related position in (Andreou 2007b). See also (McClennen 1990).

[5] This view, which figures as an easily overlooked compromise between the preceding view and the alternative line of thought I will get to presently, accords with interpreting transitivity as a precondition for the applicability of any viable theory of rational choice, and so as an essential axiom of rational choice theory, even if is is not a rational requirement on preferences.

[6] For some relevant discussion, see, for example, (Tenenbaum and Raffman 2012) and (Andreou

Figure 33.1 Read "x < y" as "y is preferred to x", and "x<y<z" as "x<y & y<z."

rationality is not in the business of evaluating an agent's basic preferences – it simply takes them as given. Second, unlike the stance that rationality can only advise an agent with respect to non-cyclic preferences, it does not have the unwelcome implication that, insofar as an agent finds herself with basic preferences that are cyclic over the available options, instrumental rationality is silent and so fails to require that the agent avoid outcomes that are unacceptable relative to her concerns when acceptable outcomes are available.

2015).

Incommensurable alternatives and gratuitous costs

The final sort of scenario that I will consider relates regret and incommensurable alternatives. I will say that two alternatives X and Y are *incommensurable* (where this evaluation might, in accordance with an instrumental conception of rationality, be relative to the agent's concerns) if, taking into account all relevant facts about the choosing agent and the circumstances of choice, X is not better than Y, Y is not better than X, and X and Y are not exactly equally good. Incommensurable alternatives are closely tied to *preference gaps*, wherein two options are not ranked in relation to one another by the agent's preferences, not even as equally desirable, and, like cyclic preferences, preference gaps can lead to a regrettable series of choices.

There is some controversy concerning whether options can really be incommensurable. On the face of it, it seems plausible to suppose that if X is not better than Y and Y is not better than X, then X and Y must be exactly equally good. But the *small-improvement argument* poses a challenge to this standard position.[1] According to the small-improvement argument, there can be situations in which

> (1) X is not better than Y,
>
> (2) Y is not better than X, and
>
> (3) a slightly improved version of X, call it X+, is better than X but not better than Y.

Since you'd think that if X and Y were exactly equally good, anything better than X would be better than Y, too, we have, it seems, good reason to think that X and Y are not exactly equally good. And plausible candidates for X, Y, and X+ have been proposed. Consider, for example, the case in which X stands for a particular cup of tea, Y stands for a particular cup of coffee, and X+ stands for a cup of tea that is slightly more fragrant than X (Chang 2002). If all that matters in the case at hand is how the beverages taste to the agent, and the agent has "authority over which [beverage] tastes better to [her]" (669), it seems possible that (1), (2), and (3) could all hold, and that X and Y could thus qualify as incommensurable.

[1] For some relevant discussions, see, for example, (de Sousa 1974), (Raz 1986, chapter 13), and (Chang 1997).

If incommensurable options exist, then some pairs of options will count as either *on a par* or else as completely *incomparable*.[2] Roughly speaking, two options count as on a par when they are not precisely comparable but are roughly equally good or in the same league. The tea and coffee case just considered is arguably a case of two options being on a par. Two options count as incomparable when they are incommensurable (in the sense identified at the beginning of this section) and cannot be compared at all, not even roughly. Perhaps, given its familiar philosophical construal as a case of incommensurability, Abraham's choice (in the the biblical story of Abraham and Isaac) between, on the one hand, obeying God and sacrificing his son and, on the other hand, disobeying God and protecting his son is a case of incomparability. Insofar as Abraham's options are incommensurable, it seems like one might plausibly resist the suggestion that, even if not exactly equally good, the options are roughly equally good; perhaps the options are too different to be comparable even as in the same league.[3]

Whether two incommensurable options are on a par or incomparable, the agent is at risk of following a series of seemingly individually permissible steps that collectively involve her incurring a gratuitous cost. Suppose that, given all the relevant facts about me and my situation, X and Y are incommensurable for me, and the same is true of X+ and Y. Accordingly, neither X and Y nor X+ and Y are ranked in relation to one another by my preferences. Suppose further that I have X+ and am asked by a passerby whether I would be willing to part with X+ for Y. Having no preference between them, I make the trade. Later, I am asked by another passerby whether I would be willing to trade Y for X. Having no preference between them, I make the trade. Both choices seem individually permissible, and yet it seems like I can sensibly regret the transition from X+ to X.

There are a variety of interesting ways of responding to this sort of case. One familiar move is to dismiss preference gaps as rationally impermissible.[4] Perhaps a rational agent is required to fill in his preference gaps in some consistent way and then act accordingly. A related move is to claim that an agent cannot seek guidance from

[2] See Chang (2002) for a discussion of the possibility of parity and a critique of jumping directly from the small-improvement argument to the conclusion that there must be incomparable options.

[3] For some helpful discussion of the case, see (Broome 2001).

[4] This view accords with interpreting the axiom of completeness (which precludes preference gaps) as a requirement of rationality rather than as just a precondition for the applicability of a certain way of representing preferences. As with the dismissal of cyclic preferences (and, more generally, intransitive preferences) as irrational, an attraction of this view is that it allows one to hang on to the idea that rational preferences are well behaved in that, so long as the preferences remain in effect, choice that is genuinely in accordance with them cannot lead to an outcome that is worse, relative to the preferences at issue, than another option that was available. In (Andreou 2005), I argue that we can hang on to both the idea that preference gaps can reflect genuinely incommensurable options and the idea that the completeness assumption is a rational requirement on preferences by interpreting the completeness assumption as concerned with "rankings settled on *for the purposes of choice*." I am now inclined to back off the strong claim that rationality requires that incommensurable options be ranked for the purposes of choice and to retreat to the weaker claim that rationality requires that an agent with preference gaps proceed with care. and, in particular, in accordance with a constraint that I will get to shortly.

rationality until he has filled in any preference gaps that are in play in some consistent way.[5]

Alternatively, it might be suggested that it is permissible to have preference gaps and that an agent with preference gaps can avoid the pitfall under consideration, so long as he proceeds with care. According to one line of thought in this ballpark, there is a rational prohibition against "brute shuffling" and respecting this prohibition protects agents with preference gaps from finding themselves in the concerning predicament described above.[6] Roughly put, in cases of brute shuffling, one settles on an option but then switches for no reason (and, in particular, without the switch being anticipated by a defensible plan to sample different options). Suppose such shuffling really is prohibited. It follows that, once an agent settles on an option, then, even if his initial choice was somewhat arbitrary because there were other equally good or else incomparable options available, there is rational pressure on him to remain with that option so long as no better option presents itself (assuming no defensible "sampling plan" was adopted); so, other things equal, a rational agent will not have to worry about shuffling to a worse version of an option he shuffled away from before.

Significantly, even among those who allow for incommensurable alternatives, there is some controversy concerning whether there is a rational prohibition against brute shuffling. For some (including myself), certain instances of brute shuffling are permissible, so long as the agent avoids "self-defeating" behavior by, for example, tracking things like whether a particular instance of brute shuffling would qualify as a shuffle to a worse version of an option she shuffled away from before.[7] According to an especially permissive view, brute shuffling need not be irrational even in cases where some gratuitous cost is incurred, though it might be if it one's shuffling is repeated enough to significantly compromise one's prospects in life by, for example, leading one to "poverty without realizing any other value or end" (Tenenbaum 2014, 403). In a variation in this view, it might be suggested that, insofar as all the options at stake are in the same league, the gratuitous costs that might be incurred by brute shuffling are relatively insignificant; the cases where shuffling must be limited are cases in which the agent gradually risks crossing, over a series of shuffles, the vague boundary between two leagues and going from one option to a significantly worse option. In such cases, there seems to be some cause for serious concern, even if one zooms out and focuses on the big picture. As in the case of cyclic preferences, an agent in such a case might need to rely on a sort of plan-centrism that, in other cases, might seem overly rigid, or else on habits that prompt her to effectively use her discretion and show restraint in good time.

[5] I defend a view in this ballpark in (Andreou 2005). As indicated in note 16, my view has evolved since then.

[6] Bratman suggests that this is at least the case for agents with an interest in self-governance (2012). The quoted phrases in this paragraph are from (Bratman 2012, 81).

[7] See, for example, (Andreou 2012) and, relatedly, (Andreou 2005).

Conclusion

I have focused on three sorts of cases in which an agent can easily find herself making a choice or series of choices that she later looks back on with regret. The cases raise some extremely interesting philosophical issues and, in each case, there is continuing lively debate concerning whether or how the agent is going wrong. One thing that is clear is that, even an agent who knows the consequences of each of her options at each choice point can make a series of choices that she would not be willing to choose as a package. This is concerning, since it is often satisfaction with one's choices put together rather than with each choice as it is made that seems crucial. It might be that, with enough constraints on one's preferences, one need not worry about proceeding piecemeal. The constraints are not, however, ones that our preferences automatically conform to, and getting them to conform, even if possible, may not be as efficient as proceeding holistically instead.[1]

[1] I am grateful to the editors for helpful comments on an earlier draft of this chapter.

References

Ainslie, George. 2001. *Breakdown of Will*. Cambridge: Cambridge University Press.

Andreou, Chrisoula. 2016. "Dynamic Choice," in *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. Fall 2016 ed., https://plato.stanford.edu/archives/fall2016/entries/dynamic-choice/.

———. 2015. "The Real Puzzle of the Self-Torturer: Uncovering a New Dimension of Instrumental Rationality," *Canadian Journal of Philosophy* 45: 562–575.

———. 2014. "Temptation, Resolutions, and Regret," *Inquiry* 57: 275–292.

———. 2012. "Self-Defeating Self-Governance," *Philosophical Issues* 22: 20–34.

———. 2007a. "Understanding Procrastination," *Journal for the Theory of Social Behaviour* 37: 183–193.

———. 2007b. "There Are Preferences and Then There Are Preferences," in *Economics and the Mind*, edited by Barbara Montero and Mark D. White. New York: Routledge.

———. 2005. "Incommensurable Alternatives and Rational Choice," Ratio 18: 249–261.

Bratman, Michael. 2014. "Temptation and the Agent's Standpoint," *Inquiry* 57: 293–310.

———. 2012. "Time, Rationality, and Self-Governance," *Philosophical Issues* 22: 73–88.

Broome, John. 2001. "Are Intentions Reasons? And How Should We Cope with Incommensurable Values?" In *Practical Rationality and Preference*, edited by Christopher W. Morris and Arthur Ripstein. Cambridge: Cambridge University Press.

Chang, Ruth. 2002. "The Possibility of Parity," *Ethics* 112: 659–688.

———. 1997. "Introduction," in *Incommensurability, Incomparability, and Practical Reason*, edited by Ruth Chang. Cambridge: Harvard University Press.

Davidson, Donald, McKinsey, J. C. C., and Suppes, Patrick. 1955. "Outlines of a Formal Theory of Value, I," *Philosophy of Science* 22: 140–160.

De Sousa, Ronald B. 1974. "The Good and the True," *Mind* 83: 534–551.

Foot, Philippa. 2001. *Natural Goodness*. Oxford: Clarendon Press.

Holton, Richard. 2009. *Willing, Wanting, Waiting*. Oxford: Clarendon Press.

Hume, David. 1978 [1739–40]. *A Treatise of Human Nature*. Edited by L. A. Selby-Bigge and P. H. Nidditch, 2nd ed. Oxford: Clarendon Press.

Korsgaard, Christine. 1996. *The Sources of Normativity*. Cambridge: Cambridge University Press.

McClennen, Edward F. 1990. *Rationality and Dynamic Choice*. New York: Cambridge University Press.

Raz, Joseph. 1986. *The Morality of Freedom*. Oxford: Clarendon Press.

Schick, Frederic. 1986. "Dutch Bookies and Money Pumps," *The Journal of Philosophy* 83: 112–119.

Smith, Michael. 1994. *The Moral Problem*. Oxford: Blackwell Publishers.

Tenenbaum, Sergio. 2014. "Minimalism About Intention: A Modest Defense," *Inquiry* 57: 384–411.

———. In press. *Rational Powers in Action*. New York: Oxford University Press.

Tenenbaum, Sergio, and Raffman, Diana. 2012. "Vague Projects and the Puzzle of the Self-Torturer," *Ethics* 123: 86–112.

Williams, Bernard. 1979. "Internal and External Reasons," in *Rational Action: Studies in Philosophy and Social Science*, edited by R. Harrison. Cambridge: Cambridge University Press.

34. Plan rationality[1]

[1] This essay draws from (Bratman 2017, 2018a). Thanks to Facundo Alonso, Jennifer Morton, Sarah Paul, Kurt Sylvan and Steven Woodworth.

Plan rationality?

An important idea in practical philosophy is that there are norms of rationality that apply to our intentions and beliefs (Harman 1976; Bratman 1987). One example is a consistency norm according to which both intending A while believing not-A, and intending A and intending B while believing that A and B are not co-possible, are rational breakdowns.[1] A second example is a norm that enjoins coherence between intended ends and intended means: there seems to be a rational breakdown in intending E, believing that a necessary means to E is M and that one will M only if one now intends M, and yet still failing to intend M.

There are debates about how exactly to understand such norms. Perhaps they articulate local or pro tanto, rather than global or on-balance, rational pressures. (Smith 2004) Perhaps one might, for example, achieve local, but not global, rationality in intending known necessary means to an intended end even though the intention in favor of the end is not sufficiently supported by what one sees as one's relevant reasons. Further, perhaps there are special circumstances in which these norms do not apply in the standard way. An example might be a planning analogue of the paradox of the preface in which the consistency norm does not apply in the standard way (Bratman 2009a: note 7; Shpall 2016; Goldstein 2016). Finally, there are issues about scope: does the means-end coherence norm have "narrow" scope and say, roughly, that if you intend E and believe M is a necessary means to E, then there is rational pressure to intend M? (Kolodny 2005). Or does it rather have "wide" scope and say that there is rational pressure, roughly, to be such that if you intend E and believe M is a necessary means to E, then you intend M? (Broome 2013).

Here I will suppose that these are norms of local or pro tanto rationality.[2] Further, I will suppose that these are wide-scope norms that enjoin a certain organization of one's mind. One can, for example, avoid the threat of means-end incoherence by intending the cited means, but one can also avoid it by giving up the intended end. There is here a normative symmetry, though it remains possible that other rational pressures (e.g., pressures for intention stability) asymmetrically favor one of these ways of avoiding incoherence.

These norms of consistency and means-end coherence directly concern the organization of one's intentions at a given time (or during a small temporal interval), though

[1] For discussion, see (Núñez 2019), (Núñez 2020), (Yaffe 2004).

[2] This is why it is appropriate to say that these norms articulate rational pressures in the direction of global, on-balance rationality.

one may well adjust one's intentions over time (e.g., by adding an intention in favor of necessary means) in order to avoid violating these norms. There is also room for norms that directly concern the development of one's intentions over time. An example might be a norm that favors a default persistence of prior intention. I will turn to this idea subsequently. For now, let's focus on the cited synchronic norms.

Consider two reasons such norms can seem puzzling. First, there are other practical attitudes with respect to which it is less plausible that analogous norms apply. Consider ordinary desires. It is a common feature of our lives that we desire things we know not to be co-possible. And it is unclear why there is any rational breakdown in desiring, for example, to be very wealthy while finding unattractive, and so not desiring, what one thinks are necessary means. If our thoughts about synchronic rationality of intention are on the right track, there must be something about intention that distinguishes it from ordinary desires and helps explain these differences in relevant rationality norms.

Second, when we reflect on these wide-scope norms, they can seem simply to be recommending a kind of mental tidiness. But why should such tidiness be a big deal? (Raz 2005; Kolodny 2008; McCann 1991).

To respond to these challenges, we need a systematic account of intention itself: there is here a tight connection between questions in normative philosophy about practical rationality and questions in the philosophy of action about the role of intention in our human agency.

My proposal highlights an apparently basic feature of our human agency: we are planning agents. Given needs for coordination over time and with each other, and given our resource limits (Simon 1955), we normally settle in advance on prior partial plans that frame further practical thinking in which we fill in these plans as need be and as time goes by. Not all purposive agents are planning agents: there can be purposive agents – sheep, perhaps – who act in order to pursue their goals and in light of their cognitive grasp of the world but who are not planning agents.[3] In contrast, we are planning agents, and this is a basic element in our capacities for intrapersonal and interpersonal organization of our agency (Bratman 1987, 2014, 2018a). And my proposal is that we understand the intentions that are the target of the cited rationality norms as states in this planning system: intentions are plan states.

To spell this out, we need a model of the normal functioning of such a planning system. Here a basic idea is that this functioning involves at least implicit guidance by the cited norms (Gibbard 1990). Given our resource limits, our prior plans will exhibit a characteristic partiality. Given this partiality, it is because of pressures for means-end coherence that one normally is led to questions of means and preliminary steps. And in solving these problems of means-end coherence, one is constrained by pressures of consistency of intention. In this normal functioning of the planning system, one's practical thinking is framed by one's prior, partial plans and is guided by one's at least implicit acceptance of norms of plan coherence and consistency.

[3] This is in the spirit of Grice's strategy of creature construction (Grice 1974). See also (Velleman

Within this model, these norms apply to intentions in a way in which they do not apply to ordinary desires. But how we can justify this distinctive normative guidance?

An initial proposal is that guidance by pressures for plan consistency and coherence supports important forms of intrapersonal and interpersonal coordination and organization. Intending what one takes to be incompatible options would normally lead one to trip over oneself, and intending ends without intending believed necessary means would normally block the efficacy of one's planning system. At least, this is true if one's relevant beliefs are reliable. And in each case, we should expect analogous impacts on relevant interpersonal coordination. It is in part because the planning system is guided by these norms that it tends to be successful in these coordination roles (Bratman 1987; Morton 2017).

However, while this is true and important, it does not suffice. Perhaps in a special case, one might better achieve one's ends by way of intentions and plans that are, given one's beliefs, inconsistent.[4] Or perhaps in another special case, one would be rewarded for means-end incoherent plans. Inconsistency or incoherence of intention in such cases still seems a kind of rational breakdown. Why? Appeal to the benefits of the general system of plan-infused practical thinking provides a defense for this *general* mode of thinking, but it is not clear how to get from a defense of this general mode of thinking to a defense of its application in each *particular* case (Smart 1956).

One response would turn to theoretical rationality. If intentions involve associated beliefs then perhaps we can see norms of plan consistency and coherence as riding piggy-back on theoretical norms of consistency and coherence of belief (Harman 1976; Setiya 2007; Wallace 2001; Velleman 2014). This is *cognitivism* about these norms of plan rationality (Bratman 1991). Such cogntivism does not yet justify the cited theoretical norms. But it is sometimes progress to reduce two problems to one.

How might this work? Suppose – for the sake of argument – that intending that p necessarily involves believing that p.[5] We could then argue that inconsistent intentions necessarily involve inconsistent beliefs. And it might seem that we could also argue that means-end incoherence of intention necessarily involves incoherence of belief. After all, if you intend E, and so (on present assumptions) believe E, and if you believe that E only if M and that M only if you intend M, and yet you do not intend M, then it may seem that you will believe that E without believing that a necessary condition for E will obtain. And that would be a kind of theoretical incoherence of belief.

But I do not think that this works, since one might believe that one intends M and yet not in fact intend M (Bratman 2009b). In such a case, one's beliefs might be coherent even though one's plans remain means-end incoherent. Granted, in such a case one has a false belief about oneself, but false belief need not be irrationality or incoherence. So, there can be means-end incoherence in one's plans without incoherence

2000: chap. 1), (Bratman 2000b).

[4] See the video games example in (Bratman 1987: chap. 8).

[5] I challenge this in (Bratman 1987: 37–39). But I think that even given this strong connection between intention and belief, cognitivism does not work.

in one's beliefs. So, this cognitivist strategy does not fully explain why means-end coherence of plans matters.

Another strategy might be to highlight a kind of inescapability.[6] After all, we have supposed that guidance by these norms is built into our planning agency. But the problem here is that, as noted, there are multiple forms of agency; in particular, one can be a goal-directed but non-planning agent. So, guidance by these norms is not inescapable for agency. So, it is not clear how the idea that guidance by these norms is partly constitutive of, in particular, planning agency can fully explain why these are norms of practical rationality.[7]

Donald Davidson highlights a different kind of inescapability. Treating norms of rationality as a single over-all package, Davidson writes that

> it is only by interpreting a creature as largely in accord with these principles that we can intelligibly attribute propositional attitudes to it... . An agent cannot fail to comport most of the time with the basic norms of rationality.
>
> (2004: 196–197)

This is a plausible idea about interpretation, but it does not solve our problem about plan rationality. What is claimed to be inescapable for a person with a mind is only failing to "comport most of the time" with relevant norms (Kolodny 2008). So we do not yet have an explanation of why a violation of these norms is, in each particular case, a rational breakdown. Further, Davidson is here treating norms of rationality as a single package, one involved quite generally in interpreting minds. But given the multiplicity of agency, there can be minded agents who are not planning agents and so are not subject to norms of, in particular, plan rationality.

What do we need to show to defend the claim that these are norms of practical rationality for a planning agent? Suppose you are – as there is pragmatic reason to be – a planning agent. Your practical thinking involves guidance by the cited norms. You can nevertheless step back and ask whether this structure of thinking makes sense. And your answer will have implications for our descriptive and explanatory model of human agency. After all, our confidence in the descriptive and explanatory significance of planning structures in human agency would be to some extent challenged if these structures would not themselves be stable under a planning agent's reflection.

So, let's focus on you – a planning agent reflecting on her characteristic forms of practical thinking and asking whether they make sense. Part of your answer will return you to the general pragmatic benefits – especially given your resource limits and needs for coordination, both cross-temporal and social – of planning agency. But you will also need to supplement these observations in order fully to defend the application of these norms to each particular case. How?

---

[6] For appeal to these forms of inescapability, see (Korsgaard 2009: 68, 82–83), (Velleman 2000).

[7] (Setiya 2014: 74–76) poses this as an issue about "pluralistic rationalism". And see (Enoch 2006).

The strategy of self-governance

Here I think we make progress by understanding the ways in which these norms track central conditions of a planning agent's self-governance. Call this the *strategy of self-governance* (Bratman 2009a, 2010, 2017, 2018a). To successfully pursue this strategy, you need to do two things: (1) Articulate ways in which these norms track conditions of a planning agent's self-governance. (2) Explain how this provides further support for these norms.

Beginning with (1), we can draw on the Frankfurt-inspired idea that an adequate model of self-governance at a time (or during a small temporal interval) involves the idea of "where (if anywhere) the person himself stands" (Frankfurt 1988: 166). In self-governance, one's thought and action are guided by where one relevantly stands, and one's standpoint is constituted by a web of relevant attitudes. To play this role, this attitudinal web needs to be sufficiently coherent (Bratman 2009a) – though such coherence does not require that the agent have, in particular, the end of synchronic self-governance. And the idea is that guidance by such a coherent web of *attitudes* can constitute the *agent's* self-governance.

The next idea is that, given the basic roles plan states play in your practical thought and action, your relevant attitudinal webs will be plan infused. In planning, one is settled on certain courses of action and/or the relevance of certain considerations to one's ongoing practical thinking.[1] One's plan states will normally cross-refer: one's plans for the future will at least potentially refer to one's future intentions, which will normally refer back to one's earlier plans. Further, there will normally be a kind of interdependence between, on the one hand, intentions at a given time and, on the other hand, past intentions and expectations of future intentions. These cross-referring, issue-settling, interdependent plans will frame one's ongoing thought and action in ways that tend to track and to support mesh between sub-plans at different times. In these multiple ways, these plan states will normally induce forms of psychological continuity and referential connectedness familiar from Lockean models of personal identity (Parfit 1984; Yaffe 2000; Bratman 2000a).

This supports the conclusion that a planning agent's self-governance at a time will involve relevant consistency and means-end coherence of plan. After all, if you intend A and intend B, but believe that A and B are not co-possible, there will be no clear answer to the question of where you stand with respect to these options. And if you intend E but nevertheless fail to intend known necessary means, even though you know

[1] The latter involves what I have called self-governing policies. (Bratman 2004)

that such an intention is now needed if you are to pursue those necessary means, then there will be no clear answer to the question of where you stand with respect to E. So, we can conclude that the cited norms of plan rationality do indeed track basic conditions of a planning agent's synchronic self-governance.

But how does this connection to self-governance provide the support for these synchronic planning norms that our reflective planning agent is seeking, support that supplements the pragmatic reasons that favor relevant general forms of thinking?

An initial answer has two prongs. First, this connection between synchronic planning norms and synchronic self-governance reveals an overarching order and commonality across those planning norms. These norms do not just track disparate forms of mental tidiness: they track conditions of self-governance. And this commonality helps make sense of these norms.

The second prong involves the idea of a normative practical reason. Suppose one has a normative reason in favor of one's self-governance. And suppose one has the capacity for relevant self-governance. Given the way in which the cited norms track necessary conditions of a planning agent's self-governance, we can plausibly infer that one will have a normative reason of self-governance to conform to these norms.[2] And the idea is that appeal to this reason is part of a two-pronged argument, available to a reflective planning agent, that in tandem with pragmatic support for general structures of plan-infused thinking supports the conclusion that these are indeed norms of practical rationality for her.[3]

To fill this in, we will need to say more about the relevant idea of a normative practical reason. But first let's ask how this approach to synchronic plan rationality might bear on the question of whether there is, as well, a norm of diachronic plan rationality. Can we extend the strategy of self-governance in the direction of a norm concerning the default stability of intention?

[2] (Bratman 2009a). For complexities concerning this inference, see (Kolodny 2018).

[3] Concerning the distinction between establishing a reason for conformity to a norm and establishing that it is a norm of rationality, see (Setiya 2014).

Extending the strategy of self-governance: 'acting together with oneself' over time

In the background are considerations concerning intention stability other than ones that appeal to self-governance. There may well be pragmatic grounds for a general tendency toward sticking with one's prior intentions. After all, this tendency may make one both more likely to resist sudden and temporary shifts in desire and a more reliable partner in joint activities. But this is at most an argument for a general disposition of thought and action, and there will in any case be a presumption against stability of prior intention in the face of new information that the agent sees as undermining the case for the intended action.

To this we can add two ideas (Bratman 1987). First, a prior intention at t_1 to A at t_3 may well lead to action that changes the downstream circumstances in ways that help support continuing to intend at t_2 to A at t_3. This is the snowball effect. Second – and especially for resource-limited agents like us – reconsideration of a previously formed intention involves characteristic costs and risks to previously forged coordination. So, in many cases, it will be sensible not to reconsider. And such sensible non-reconsideration will support the stability of one's prior intention.

Against this background, the extended self-governance strategy helps us articulate a further consideration in favor of stability of intention. It does this by way of appealing not only to self-governance *at* a time but also to self-governance *over* time.

What is a planning agent's self-governance *over* time? My proposal is that a planning agent's self-governance over time involves her self-governance at times along the way together with relevant cross-temporal interconnections between these instances of synchronic self-governance. (Though see Nefsky and Tenenbaum forthcoming.) What interconnections? Here I propose a metaphor: in governing her own activity over time, a planning agent is 'acting together with herself' over time (Bratman 2018b). This will involve both self-governance at times along the way and intrapersonal cross-temporal interconnections that are analogous to the *inter*personal interconnections of plan states that are characteristic of interpersonal shared intentional activity, as understood in Bratman (2014). These intrapersonal interconnections will include characteristic forms

of intrapersonal continuity, cross-reference, interdependence, and mesh of intention over time.[1]

If that is how we understand a planning agent's self-governance over time, might we see diachronic plan rationality as, in part, tracking such diachronic self-governance? Well, if it did, then we would have the basis for a (limited) presumption in favor of continuity of intention over time, so long as that continuity coheres with self-governance at times along the way. This would be a kind of conservatism, one that gives a prior intention a (limited) default status.[2] But it would be only a modest conservatism, since this default status would be present only so long as following through with one's prior intention itself cohered with synchronic self-governance at the time of follow through.

[1] Given the hierarchical structure of plans, there can be such interconnections at a higher level despite the absence of such interconnections at a lower level.

We can in this way construct diachronic self-governance out of synchronic self-governance and relevant interconnections while still maintaining that the self-governance that is of primary interest to us is diachronic.

[2] For alternative approaches, see (Ferrero 2012), (Paul 2014).

Shuffling

Consider now a case in which one decides on a temporally extended option in the face of what one sees as non-comparable considerations in favor of conflicting options (Broome 2001; Bratman 2012). In a version of Sartre's example, one decides in favor of staying with one's mother rather than fighting with the Free French (Sartre 1975). Given that the relevant activity is extended over time, the non-comparability will normally remain as time goes by. Is there nevertheless rational pressure to stick with one's earlier decision rather than to shuffle to a decision in favor of the alternative?

The proposed connection between diachronic plan rationality and diachronic self-governance points to an affirmative answer. Given that the non-comparability continues to be recognized by you, sticking with your prior intention to stay with mother, and switching to an intention in favor of the Free French, would each, taken separately, cohere with your then present evaluation. But if you stick with your prior intention your relevant intentions over time will exhibit a constancy characteristic of a planning agent's self-governance over time; if, in contrast, you change your mind in favor of the Free French, there will be a kind of intention discontinuity that is in tension with self-governance over time. So, conditions of diachronic self-governance favor the stability of your prior intention. So, if diachronic plan rationality tracks conditions of diachronic self-governance, it will in this respect favor this stability.

Temptation

Consider now a puzzle about rational willpower (Holton 2009; Bratman 1998). You know you will be tempted to drink a lot at the party. But you think that would be a mistake. So, you now decide in advance to have only one drink. You know, however, that at the party, your judgment will shift, and you will newly judge it best to have many drinks. So, if you were nevertheless to follow through with your prior intention, you would be acting contrary to your then-present evaluation. And this is so even though, as you also know, you would even later regret having given into temptation. So, how could it be rational for you to follow through on your prior intention?

Can our comments about a planning agent's diachronic self-governance help here? Well, if you follow through with your prior intention, rather than giving into temptation, there will be continuity in intention. So, if diachronic plan rationality tracks conditions of a planning agent's diachronic self-governance it may seem we have a way of explaining how such willpower can sometimes be rational.

The problem is that continuity of prior intention helps constitute diachronic self-governance only if there is synchronic self-governance at times along the way. But given your shift in evaluative judgment, it seems that your stance at the time of the party favors acting contrary to your prior intention. So, in following through with your prior intention, you will not be synchronically self-governed.

What we need, I think, is an end of the agent's that in some way favors relevant intention continuity and thereby can potentially help re-shift her standpoint at the time of the party so that it supports intention follow-through. Such an end would open up the possibility that following through with one's prior intention in a temptation case is indeed a case of synchronic self-governance and so, since it involves relevant intention continuity, a case of diachronic self-governance.

What end? A simple appeal to an end of intention continuity would face the worry that we are just appealing to a kind of diachronic mental tidiness.[1] A different approach would appeal to an intellectual end of self-understanding (Velleman 2006: 272). But given our search for commonality in our understanding of these norms, this would lead back to the cognitivism against which I have argued. So, I propose, instead, that we appeal to the end of one's diachronic self-governance, as we have been understanding it. This end would potentially support willpower in the face of temptation, since such willpower would involve a continuity of intention that is an element in diachronic self-

[1] For appeals to an end of continuity, see (Sobel 1994), (Rabinowicz 1995).

governance.[2] Given support from this end, the agent's standpoint at the time of the party might well shift back in favor of willpower in a way needed for that willpower to be a case of synchronic self-governance.

The next point is that problems of temptation, and related problems about procrastination, pervade our human lives (Paul 2014; Tenenbaum and Raffman 2012; Andreou 2014). So, we can expect that this end of one's diachronic self-governance will be an element in central cases of the exercise of a human planning agent's capacity for diachronic self-governance. Whereas synchronic self-governance does not in general require the end of synchronic self-governance, there is pressure on the diachronic self-governance of human planning agents to involve the end of diachronic self-governance. This supports the idea that if diachronic plan rationality were to track conditions of a planning agent's diachronic self-governance, it would not only favor intention continuity, in a context of synchronic self-governance at times along the way, it would also favor the presence of the end of one's diachronic self-governance. And this end, if present, can sometimes shift what is supported by the agent's standpoint at the time of plan follow-through. The conjecture, then, is that this extension of the strategy of self-governance to diachronic plan rationality would support both a modest norm of default stability of prior intention and the presence of the end of one's diachronic self-governance.

The idea is not that this end of diachronic self-governance is essential to agency quite generally.[3] There are, as noted, multiple kinds of agents; it is not clear that young human planning agents quite generally have this end, and even synchronically self-governed agency, taken on its own, does not require an end of self-governance. The proposed connection to this end of diachronic self-governance goes instead by way of a strong form of temporally extended agency, namely: diachronically self-governed human planning agency.

[2] A more complete story would also appeal to the agent's expected later regret, if she were now to give into temptation, as a difference between such a temptation case and a version of Kavka's toxin case. (Kavka 1983; Bratman 1998).

[3] Versions of this stronger idea are in (Korsgaard 2009), (Velleman 2000).

A unified account

Thinking about diachronic plan rationality along these self-governance-based lines sets the stage for a uniform self-governance-based form of support for both the synchronic and the diachronic aspects of plan rationality, a form of support that supplements the pragmatic support for general forms of plan-infused practical thinking. Norms of synchronic and diachronic plan rationality are tied together, and thereby made more intelligible, by their property of tracking conditions of a planning agent's self-governance, both synchronic and diachronic. This then supports an extension of our earlier observation about the significance of a normative reason for self-governance. Suppose that one has a normative reason in favor of one's self-governance, both synchronic and diachronic. And suppose that one has the capacity for relevant self-governance. Given the way in which these norms track necessary conditions of a planning agent's self-governance, both synchronic and diachronic, we can conclude that one will have a normative reason of self-governance to conform to these norms.[1]

These norms of plan rationality have a kind of stringency (Bratman 1987: 24). We can interpret this by appeal to the contrast between a pro tanto and a merely prima facie norm. The norms we have been defending do not just cite prima facie evidence in favor of a conclusion about on-balance rationality, evidence that is potentially misleading. Instead, these norms each articulate a kind of pro tanto rational breakdown that would be constituted by their violation, a pro tanto breakdown that would remain even if, in special circumstances, it would make most sense, on balance, to violate the norm. And a reason of self-governance to conform to these norms in the particular case would help support this pro tanto demand.

But what is the idea of a normative practical reason that is at work here, and will there indeed be such a reason in favor of a planning agent's self-governance?

[1] In (Bratman 1987: 24–27), I note that the idea that intentions quite generally provide new reasons threatens to lead to an unacceptable "bootstrapping" of reasons. In (Bratman 2009a, 2012), I explain how the present proposal is compatible with this point.

A reason for, and end of, self-governance?

For our purposes of articulating the relevant reflections of a planning agent, it is plausible to understand such normative reasons by appeal both to the agent's ends and to the desirability of what those ends favor. A planning agent who reflects on her own practical thinking will be interested in both what is needed to realize her ends and whether what those ends favor is desirable. So, for present purposes, I will suppose that a consideration is, in the relevant sense, a normative reason for S to A only if it helps explain why S's A-ing is needed to realize relevant ends of S[1] and only if what those ends favor is desirable. On the plausible assumption that it is desirable to govern one's own life,[2] a central question about a purported reason for self-governance will then concern the status of the end of one's self-governance.

We cannot infer simply from the desirability of self-governance that a rational agent will have the end of her self-governance. There are many good things and not enough time. One thought here might be that this end of diachronic self-governance is simply a rationally optional, contingent end. If present, it can help support the reason for self-governance for which we are looking, but there is no guarantee that it will be present. At the other extreme is the thought that this end is essential to agency. However, the former view seems too weak for purposes of defending norms of plan rationality, and, for reasons noted, the latter view seems too strong.

Is there a view in the middle? A planning agent who reflects on the basic structures of her practical thinking – and so on the cited planning norms – would recognize both that there is pragmatic support for these general, plan-infused modes of thinking and that these norms have in common the property of tracking conditions of her self-governance, both synchronic and diachronic. She would go on to see that an end that is an element in the normal exercise of her capacity for diachronic self-governance – the end of her diachronic self-governance – would, if present, also induce an end of synchronic self-governance. If present, this combined end of self-governance at a time and over time would ground a normative reason for both diachronic and synchronic self-governance, and this reason would then support the application of her norms of

[1] (Williams 1981; Schroeder 2007) – though Schroeder appeals to sufficient, whereas I appeal to necessary means.

[2] Which is not to say that all self-governance – even self-governance in the pursuit of bad ends – is good on balance.

plan rationality to the particular case. Since the norm of diachronic plan rationality that would thereby be supported by this end of self-governance itself supports the presence of this end, this end of self-governance would be, if present within this planning framework, *rationally self-supporting*. Given this rationally self-supporting end – an end that favors a central, organizing commonality across her planning norms – her package of pragmatically supported plan structures and ends would be reflectively stable: this would be a *rationally stable reflective equilibrium*. So, given her end of self-governance, as well as basic pragmatic pressures, it would make sense for her to retain her plan-infused practical thinking and its associated norms, norms that support that end of self-governance. These norms would thereby have for her a thoughtful and rational stability, one that depended *inter alia* on the rationally self-supporting presence of the end of her self-governance.

The significance of a stable, rational equilibrium?

This leads to a final question: Does it suffice for our supplement to a pragmatic defense of these planning norms to show that their acceptance is an element in a rationally stable reflective equilibrium that would be characteristic of a diachronically self-governing human planning agent? Is this a philosophically adequate path between the Scylla of a merely contingent, rationally optional end of one's self-governance and the Charybdis of an insistence that this end is essential to agency? Our investigation into norms of plan rationality has led us to this general question about the philosophical significance of deep, fecund, and entrenched structures of practical thinking that are in a rationally stable reflective equilibrium though they are not strictly necessary for agency, or even planning agency, *per se*.[1]

[1] For further discussion that highlights a parallel with Strawson's (2003) discussion of our framework of reactive attitudes, see (Bratman 2018a). My proposal there is that the end of one's diachronic self-governance plays a role in our reflectively stable planning framework that is to some extent similar to the role, in Strawson's view, that the concern with quality of will plays within our framework of reactive attitudes. While in neither case is the underlying end or concern strictly essential to agency or mind, in each case, it plays a fundamental role in our human lives. So, the end of one's diachronic self-governance plays a role in our theory of plan rationality that is in a way similar to the role played by the resource limitations highlighted by (Simon 1955): in each case, there is appeal to deep but not-strictly-necessary features of our actual human planning agency. As Elijah Millgram (2019) highlights, this points in the direction of a kind of psychologism about plan rationality.

Note

Michael E. Bratman is U. G. and Abbie Birch Durfee Professor in the School of Humanities and Sciences and Professor of Philosophy at Stanford University, U.S.A. His most recent book is *Planning, Time, and Self-Governance: Essays in Practical Rationality* (2018).

References

Andreou, C. (2014) "Temptation, Resolutions, and Regret," *Inquiry* 57: 275–92.

Bratman, M.E. (1987) *Intention, Plans, and Practical Reason*, Cambridge: Harvard University Press. (Re-issued CSLI Publications 1999.)

Bratman, M.E. (1991) "Cognitivism About Practical Reason," reprinted in *Faces of Intention: Selected Essays on Intention and Agency*, New York: Cambridge University Press, 1999, 250–264.

Bratman, M.E. (1998) "Toxin, Temptation, and the Stability of Intention," reprinted in *Faces of Intention*, New York: Cambridge University Press, 1999, 58–90.

Bratman, M.E. (2000a) "Reflection, Planning, and Temporally Extended Agency," reprinted in *Structures of Agency: Essays*, New York: Oxford University Press, 2007, 21–46.

Bratman, M.E. (2000b) "Valuing and the Will," reprinted in *Structures of Agency*, New York: Oxford University Press, 2007, 47–67.

Bratman, M.E. (2004) "Three Theories of Self-Governance," reprinted in *Structures of Agency*, New York: Oxford University Press, 2007, 222–253.

Bratman, M.E. (2009a) "Intention, Practical Rationality, and Self-Governance," reprinted in *Planning, Time, and Self-Governance: Essays in Practical Rationality*, New York: Oxford University Press, 2018, 76–109.

Bratman, M.E. (2009b) "Intention, Belief, Practical, Theoretical," reprinted in *Planning, Time, and Self-Governance*, New York: Oxford University Press, 2018, 18–51.

Bratman, M.E. (2010) "Agency, Time, and Sociality," reprinted in *Planning, Time, and Self-Governance*, New York: Oxford University Press, 2018, 110–131.

Bratman, M.E. (2012) "Time, Rationality, and Self-Governance," reprinted in *Planning, Time, and Self-Governance*, New York: Oxford University Press, 2018, 132–148.

Bratman, M.E. (2014) *Shared Agency: A Planning Theory of Acting Together*, New York: Oxford University Press.

Bratman, M.E. (2017) "Rational Planning Agency," reprinted in *Planning, Time, and Self-Governance*, New York: Oxford University Press, 2018, 202–223.

Bratman, M.E. (2018a) "Introduction: The Planning Framework," in *Planning, Time, and Self-Governance*, New York: Oxford University Press, 2018, 1–17.

Bratman, M.E. (2018b) "A Planning Agent's Self-Governance Over Time," in *Planning, Time, and Self-Governance*, New York: Oxford University Press, 2018, 224–249.

Broome, J. (2001) "Are Intentions Reasons? And How Should We Cope with Incommensurable Values?" In C.W. Morris and A. Ripstein (eds.) *Practical Rationality and Preference: Essays for David Gauthier*, Cambridge: Cambridge University Press.
Broome, J. (2013) *Rationality Through Reasoning*, Hoboken, NJ: Wiley-Blackwell.
Davidson, D. (2004) "Incoherence and Irrationality," in D. Davidson (ed.) *Problems of Rationality*, Oxford: Oxford University Press.
Enoch, D. (2006) "Agency, Schmagency: Why Normativity Won't Come from What Is Constitutive of Action," *Philosophical Review* 115: 169–198.
Ferrero, L. (2012) "Diachronic Constraints of Practical Rationality," *Philosophical Issues* 22: 144–164.
Frankfurt, H. (1988) "Identification and Wholeheartedness," in H. Frankfurt (ed.) *The Importance of What We Care About*, Cambridge: Cambridge University Press.
Gibbard, A. (1990) *Wise Choices, Apt Feelings*, Cambridge: Harvard University Press.
Goldstein, S. (2016) "A Preface Paradox for Intention," *Philosophers' Imprint* 16.
Grice, P. (1974) "Method in Philosophical Psychology (From the Banal to the Bizarre)," *Proceedings and Addresses of the American Philosophical Association* 48: 23–53.
Harman, G. (1976) "Practical Reasoning," *Review of Metaphysics* 29 (3): 431–463.
Holton, R. (2009) *Willing, Wanting, Waiting*, Oxford: Oxford University Press.
Kavka, G. (1983) "The Toxin Puzzle," *Analysis* 43: 33–36.
Kolodny, N. (2005) "Why Be Rational?" *Mind* 114: 509–563.
Kolodny, N. (2008) "The Myth of Practical Consistency," *European Journal of Philosophy* 16: 366–402.
Kolodny, N. (2018) "Instrumental Reasons," in D. Star (ed.) *The Oxford Handbook of Reasons and Normativity*, Oxford: Oxford University Press.
Korsgaard, C.M. (2009) *Self-Constitution: Agency, Identity, and Integrity*, Oxford: Oxford University Press.
McCann, H. (1991) "Settled Objectives and Rational Constraints," *American Philosophical Quarterly* 28: 25–36.
Millgram, E. (2019) "Review of *Planning, Time, and Self-Governance*," in *Notre Dame Philosophical Reviews*, New York: Oxford University Press, May 15.
Morton, J. (2017) "Reasoning Under Scarcity," *Australasian Journal of Philosophy* 95: 543–559.
Nefsky, J. and Tenenbaum, S. (forthcoming) "Extended Agency and the Problem of Diachronic Autonomy," in C. Bagnoli (ed.) *Time in Action*, New York: Routledge.
Núñez, C. (2019) "Requirements of Intention in Light of Belief," *Philosophical Studies*. https://doi.org/10.1007/s11098-019-01321-0.
Núñez, C. (2020) "An Alternative Norm of Intention Consistency," *Thought: A Journal of Philosophy*. DOI: 10.1002/tht3.453.
Parfit, D. (1984) *Reasons and Persons*, Oxford: Clarendon Press.
Paul, S. (2014) "Diachronic Incontinence Is a Problem in Moral Philosophy," *Inquiry* 57: 337–355.

Rabinowicz, W. (1995) "To Have One's Cake and Eat It Too: Sequential Choice and Expected-Utility Violations," *Journal of Philosophy* 92: 586–620.
Raz, J. (2005) "The Myth of Instrumental Rationality," *Journal of Ethics and Social Philosophy* 1: 1.
Sartre, J.P. (1975) "Existentialism Is a Humanism," in W. Kaufmann (ed.) *Existentialism from Dostoevsky to Sartre* (rev. and expanded), New York: Meridian/Penguin.
Schroeder, M. (2007) *Slaves of the Passions*, Oxford: Oxford University Press.
Setiya, K. (2007) "Cognitivism About Instrumental Reason," *Ethics* 117: 649–673.
Setiya, K. (2014) "Intention, Plans, and Ethical Rationalism," in M. Vargas and G. Yaffe (eds.) *Rational and Social Agency: The Philosophy of Michael Bratman*, New York: Oxford University Press.
Shpall, S. (2016) "The Calendar Paradox," *Philosophical Studies* 173: 801–825.
Simon, H. (1955) "A Behavioral Model of Rational Choice," *The Quarterly Journal of Economics* 69 (1): 99–118.
Smart, J.J.C. (1956) "Extreme and Restricted Utilitarianism," *Philosophical Quarterly* 6: 344–354.
Smith, M. (2004) "The Structure of Orthonomy," in J. Hyman and H. Steward (eds.) *Agency and Action*, Cambridge: Cambridge University Press.
Sobel, J.H. (1994) "Useful Intentions," in J.H. Sobel (ed.) *Taking Chances: Essays on Rational Choice*, Cambridge: Cambridge University Press.
Strawson, P. (2003) "Freedom and Resentment," in G. Watson (ed.) *Free Will* (2nd ed.), Oxford: Oxford University Press.
Tenenbaum, S. and Raffman, D. (2012) "Vague Projects and the Puzzle of the Self-Torturer," *Ethics* 123: 86–112.
Velleman, J.D. (2000) *The Possibility of Practical Reason*, Oxford: Oxford University Press.
Velleman, J.D. (2006) "The Centered Self," in J.D. Velleman (ed.) *Self to Self*, Cambridge: Cambridge University Press.
Velleman, J.D. (2014) "What Good Is a Will?" in M. Vargas and G. Yaffe (eds.) *Rational and Social Agency: The Philosophy of Michael Bratman*, New York: Oxford University Press.
Wallace, R.J. (2001) "Normativity, Commitment, and Instrumental Reason," *Philosophers' Imprint* 1 (3): 1–26.
Williams, B. (1981) "Internal and External Reasons," in *Moral Luck*, Cambridge: Cambridge University Press.
Yaffe, G. (2000). *Liberty Worth the Name: Locke on Free Agency*, Princeton, NJ: Princeton University Press.
Yaffe, G. (2004) "Trying, Intending and Attempted Crimes," *Philosophical Topics* 32: 505–532.

35. Between sophistication and resolution – wise choice

Wlodek Rabinowicz

An action plan specifies a sequence of actions to be taken by an agent. If necessary, it contains instructions as to how the agent is to respond to possible contingencies. The agent is *dynamically inconsistent* if she embarks upon a plan but then at some point deviates from it. Actually, this is only one form of dynamic inconsistency. Even in the absence of planning, people can be dynamically inconsistent if their actions over time 'contradict' each other in various ways – if what they achieve by earlier actions is undone by their subsequent behavior. Indeed, the absence of planning increases the probability of such behavioral inconsistencies. In this chapter, though, I will focus on plan deviations and, in particular, on ways to avoid them.

Deviations from adopted plans might depend on the agent ceasing to act in accordance with her preferences, on her preferences not being wholly in line with the plan she has adopted, or on a change in her preferences. Indeed, preference change might be endogenous to the plan implementation: it might be caused by the very actions the agent performs following her plan. We are all familiar with cases like this. At a party, the agent plans to have just a couple of drinks and then call it quits. But the drinks she takes cause her to lose her restraint and continue to drink copiously.

In this chapter, however, I am primarily interested in dynamic inconsistencies that do *not* depend on such distorting changes in the agent's preferences. Instead, I will focus on the threat of dynamic inconsitency that faces an agent whose preferences violate expected utility axioms.[1]

Thus, for example, such an agent's preferences might be cyclic: She prefers X_1 to X_2, X_2 to X_3, ... , X_{n-1} to X_n, and X_n to X_1. Or, they might violate contraction consistency: an outcome she most prefers in a larger set is not most preferred by her in a subset. Or, her preferences might violate the Independence Axiom: she would rather have a lottery in which one of the possible outcomes is replaced by an outcome she prefers less. A violation of Independence is exemplified by the famous Allais problem (Allais 1953). The problem arises for agents who put a premium on safety: they prefer safe bets to risky ones even if the risk is small and the potential gain from taking the risk is large. However, if safety is not an option and all the available bets are risky, they are willing to take higher risks for the sake of larger gains. I am going to focus on a version of this problem as my prime example in what follows (Section 1).

In all such cases, the agent can be confronted with sequential decision problems in which she might easily slide into dynamic inconsistency. Thus, to give an example,

[1] Another type of case in which dynamic inconsistency might arise without preference change is one in which the agent's discount rate for time is hyperbolic rather than exponential. Say she originally doesn't view some future pleasure to be worth the cost, but then, as the pleasure comes nearer, its utility hyperbolically increases, which makes the cost acceptable (Ainslie 1992). But doesn't this mean that the agent's preferences change? I don't think so. It is arguable that time-discounting does not involve preference change provided that the objects of preferences are taken to include time distance: "X in a month," "X in an hour," and so on. Indeed, the same effect might arise even with exponential discount rates, provided these can differ for different goods and bads, for example, for pleasures and their costs.

an agent who violates contraction consistency might originally plan to achieve her most preferred outcome X, but then, when as a result of her actions towards that end some of the other originally available outcomes drop out, she comes to face a smaller, contracted outcome set. In this set, though it still contains X, X is not her most preferred outcome, which motivates her to deviate from the originally adopted plan. Such inconsistencies might be contrary to the agent's interests; indeed, in some decision problems, they might be to her guaranteed disadvantage. Dramatically put, the agent who violates the expected utility axioms might be exploited by a clever opponent.

Vulnerability to exploitation has been used in pragmatic arguments for these axioms and for other rationality constraints.[2] Preferences that violate the axioms are thought to be irrational *because* they are 'self-defeating': they make the agent perform badly in their own terms. It is a matter of dispute, however, how compelling such arguments are.[3]

In this chapter, though, my focus is on a different issue. I am interested in *policies* of decision-making that prevent dynamic inconsistency even if the agent happens to violate expected utility axioms. While different, this issue is indirectly relevant to the status of pragmatic arguments. If dynamic inconsistency is avoidable given appropriate decision policies, then pragmatic justification of expected utility axioms becomes a much harder task.

There are two well-known candidates for policies of this kind: *sophisticated* and *resolute* choice. The former involves looking ahead and reasoning backwards. The agent identifies the best move at the ultimate choice node on each branch of the decision tree, predicts that she is going to make this move if she will reach the node in question, and then, using that prediction, determines the best move at the penultimate choice node on each branch. Continuing in this way, she determines the best moves at all choice nodes, including the first one. We shall see in Sections 2 and 3 how sophisticated choice works in more detail. Resolute choice, on the other hand, is based on the idea that a well-functioning agent is fully committed to the action plans she follows: her commitment to the chosen plan makes her act accordingly, without deviations. (Cf. Section 4.)

Resolute choice satisfies *Reduction to Normal Form*: It implies that whenever a plan is optimal in a reduced, 'one-shot' form of the decision problem, it is also optimal in its original sequential form. This condition is violated by sophisticated choice, which instead satisfies the condition of diachronic *Separability* that is violated by resoluteness.

[2] See, for example, Skyrms (1993) and Rabinowicz (2000), (2008). There are also pragmatic arguments to the effect that violators of relevant constraints miss the opportunity of receiving costless benefits. Cf. Rabinowicz (2001).

[3] One issue is whether the arguments in question really manage to establish that violations of the expected utility axioms indeed make the agent vulnerable to exploitation. But even if they do, what does it show? Is dynamic consistency required by rationality? That it is has been argued by Bratman (2012, 2014), among others. For the opposing view, see Paul (2014).

Roughly, Separability states that what is rational at a choice node in a decision tree never depends on what has taken place up to that point: it is only the future, but not the past, that matters.

Arguably, neither Separability nor Reduction to Normal Form are intuitively compelling conditions. Both of them are rejected in an alternative approach to dynamic decision making developed in Rabinowicz (1995, 1997). Rather demagogically, I called it Wise Choice. I will describe it in Section 5, but then, in Section 6, I will consider to what extent it represents a new perspective on sequential decision problems. More precisely, I will consider whether wise choice can be reduced to an appropriately reconstructed form of sophisticated choice.[4]

[4] With the exception of this last section, the chapter draws upon, but also significantly revises, Rabinowicz (1997). Cf. also Rabinowicz (1995).

1 Violation of Independence and dynamic inconsistency

For any outcomes X and Z, let XpZ stand for the lottery that yields X with probability p and Z with the remaining probability $1 - p$. The axiom of Independence concerns lotteries of this kind:

> **Independence**: For all outcomes X, Y, and Z and all probabilities $p > 0$, $XpZ \boxtimes YpZ$ iff $X \boxtimes Y$.

Outcomes X, Y, and Z may themselves be lotteries over other outcomes. $\boxtimes$ stands for weak preference. Thus, $X \boxtimes Y$ should be read as: "The agent strictly prefers X to Y or is indifferent between X and Y." Strict preference, or simply "preference", as I will often write in what follows, is defined as the asymmetric part of $\boxtimes$: $X \boxtimes Y$ iff $X \boxtimes Y$ but not $Y \boxtimes X$. Indifference is the symmetric part of $\boxtimes$: $X \boxtimes Y$ iff $X \boxtimes Y$ and $Y \boxtimes X$.

Given this definition of $\boxtimes$, it follows from Independence that $XpZ \boxtimes YpZ$ if and only if $X \boxtimes Y$. In other words, improving a possible lottery outcome improves the lottery itself. To this extent then, each outcome makes an independent contribution to the value of the lottery as a whole – independent of the alternative outcome.

Many agents violate Independence in Allais-type situations. Thus, suppose X is a safe large gain, while Y is a gamble that, if won, would yield a considerably larger gain. If you play Y, you may end up with nothing, but the risk of it is slight. Z is the null outcome: you get nothing. Let probability p be neither extremely high nor extremely low. Thus, say, X = \$30,000, Y = a 0.98 chance of \$50,000, while $p = 0.5$. For these values, XpZ = a 0.5 chance of \$30,000, while YpZ = a 0.49 chance of \$50,000. ($0.5 \times 0.98 = 0.49$.) One might well prefer the safe gain X to the slightly risky Y and yet prefer YpZ to XpZ, in violation of Independence. A risk-aversive agent who would opt for the safe gain might still be willing to take a more risky gamble for the sake of a significantly larger gain, if safety is not an option and the increase in risk is small.

Consider how such preferences can get the agent to act in a dynamically inconsistent way. Let E be a random event with a known chance equal to p: $P(E) = p$. Suppose the agent also knows that E is causally independent of her actions. We assume, in addition, that apart from its effects E is of no value to the agent, either in itself or in combination with other events: it is a 'neutral' event of the type Heads or Tails. Indeed, if p = 0,5, we can let E consist in a fair coin coming up heads in the next

toss. Now, suppose the following decision tree represents the agent's sequential choice problem. Squares stand for *choice nodes* – the points at which the agent makes a move – while circles represent *chance nodes* – the points at which Nature moves. The agent learns Nature's move before making her next choice (if any).

In this example, after the agent's choice in the first node, Nature determines whether E occurs. Nature's move is independent of the agent's: It would have been the same had the agent acted otherwise. Different combinations of the agent's and Nature's moves – different branches of the tree – lead to different outcomes for the agent, specified at the end of each branch.[1] ε is a small payment that the agent has to make if she goes down in the first choice node. By contrast, going up doesn't incur any cost. If the agent goes down, it is Nature that decides the final outcome, $Y - \varepsilon$ or $Z - \varepsilon$, but if the agent goes up and E occurs, she gets to make the second choice, between X and Y. If E doesn't occur after she has gone up, the agent ends up with Z (nothing).

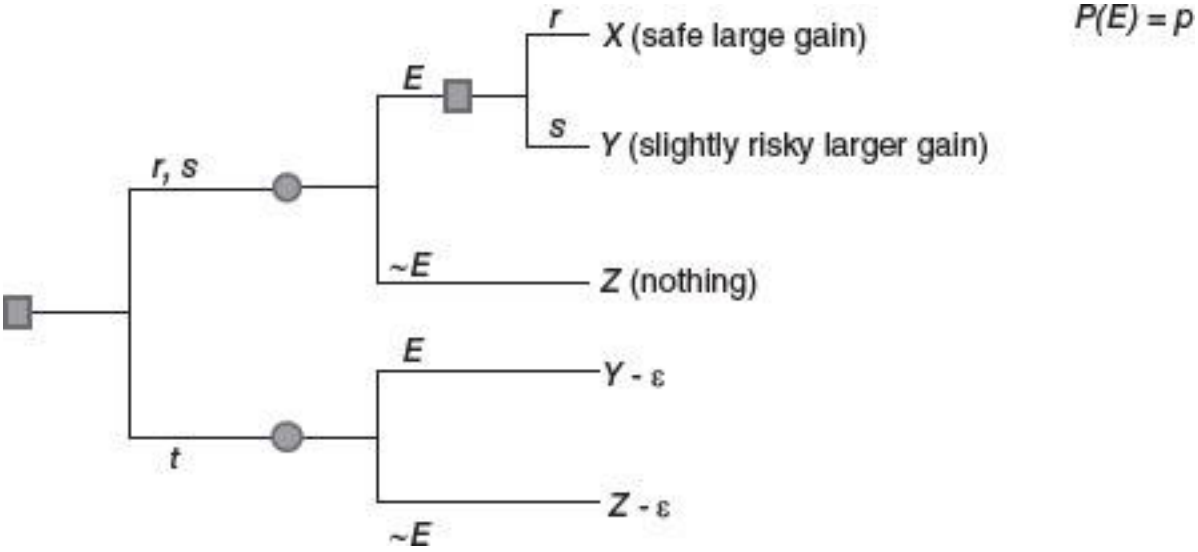


*Figure 35.1*The Allais problem in sequential form

A *plan* describes how the agent is to act at different choice nodes. It specifies her moves at all the choice nodes that she can reach, possibly with Nature's help, by making earlier moves specified by the plan. Thus, plan r instructs her to go up, and then, if E occurs, to go up again. Plan s also prescribes going up in the first node, but then going down if E occurs. Finally, plan t prescribes going down in the first node.

A plan prepares the agent for 'natural' contingencies – for the moves done by Nature – but not for her own potential deviations from the plan. For example, plan t tells the agent to go down in the first choice node but is silent as to what she should need to do if she instead were to go up in that node, contrary to what t prescribes. It doesn't specify

[1] For simplicity, I haven't drawn in Figure 35.1 Nature's second move, the one that determines the

whether she should then go up or down in the second choice node. An instruction that prepares the agent even for such eventualities and thus specifies he agent's moves in very choice node in the tree is what game theorists call a 'strategy.' Here, for simplicity, I only consider possible plans, but this doesn't substantially affect the argument that follows.

With each plan, we can associate the expected outcome of its implementation. When a plan involves chance nodes, its expected outcome is a lottery. In our example, since $P(E) = p$, we have the following expected outcomes for the different plans:

1. r ⊠ XpZ[2]
2. s ⊠ YpZ
3. t ⊠ (YpZ) – ε (i.e., ($Y - \varepsilon$)p($Z - \varepsilon$))

We know that the agent prefers YpZ to XpZ. If payment ε is sufficiently small, it doesn't reverse this preference. We may therefore assume that YpZ ⊠ (YpZ) – ε ⊠ XpZ.

If the agent chooses the plan that leads to the expected outcome she most prefers, she opts for plan s. Consequently, she goes up at the first choice node, intending to go down at the second choice node if E were to occur. But, if E does occur, the agent comes to face a choice between X and Y. Since X is the alternative she prefers, she deviates from her original plan and goes up at the second choice node. She is dynamically inconsistent.

---

outcome of lottery Y.

[2] Why have I associated XpZ with plan r and not the more complex outcome (X & E)p(Z & E)? This simplification is justified, since E has been assumed to be neutral in value, not only in itself but also in combination with other events.

2 Sophisticated choice

There is an obvious objection to this inconsistency argument: If an agent whose preferences violate Independence has foresight, she is not going to act in this dynamically inconsistent way. A provident agent predicts that she would go up at the second node and thus foresees that she would not stand by plan *s*. Indeed, she predicts that if she were to go up in the first node, this would lead to the least preferred outcome, *XpZ*, due to the expected choice of *X* at the second choice node. She therefore goes down in the first choice node, in order to secure the second best expected outcome: (*YpZ*) – ε.

This kind of provident policy is often called the *sophisticated* approach (Hammond 1988). It relies on *backward induction* – on reasoning backwards from the terminal choice nodes of the decision tree to its root. A sophisticated agent identifies the best move at the last choice node on each branch. Relying on her future rationality, she expects to make this move if she were to reach that node. Indeed, she expects to retain this expectation at the next-to-last choice node, which allows her to determine her best move at that node as well. She continues her deliberation in this manner, moving backwards to the root of the decision tree.

An action plan is *performable* from the agent's point of view if she can expect to follow it through provided she embarks upon it. Consequently, it is performable if each of her moves in the plan, apart from the one at the root of the decision tree, can be shown to be best using backward induction.[1] There are two such plans in our example: *r* and *t*. The sophisticated agent chooses a performable plan that leads to a best expected outcome as compared with other performable plans: thus, she chooses plan *t* in our example.

Advocates of the sophisticated approach take dynamic inconsistency to be due to myopia. A myopic agent fails to make (realistic) predictions about her future actions and therefore treats all the theoretically possible plans as equally performable. Because of this failure in foresight, she embarks upon a best theoretically possible plan, risking to deviate from it at a later stage

Differently put, the myopic agent treats decision problems in *extensive form*, that is, problems in which she makes choices sequentially, at more than one occasion, as though they were problems in *normal form*, in which she only needs to make a choice of

[1] What about ties? If there are several best moves at a choice node *n*, one can pick any of them as an element of a given plan and then continue backward-induction reasoning on the assumption that one is going to make this move. This assumption presupposes, though, that an agent never gives up on

an action plan, which she then automatically implements. Had our example been of the latter type, plan *s* would have been performable. No occasion for dynamic inconsistency would arise, since the agent would have had just one choice to make – to set a plan in motion.

Thus, the sophisticated approach, which takes seriously the possibility that the agent might have reasons to deviate from her plan, however good it looks from the outset, rejects the following assumption that is implicit in the myopic choice:

> **Reduction to Normal Form**: An action plan is optimal in a decision problem in extensive form if and only if it is optimal in the corresponding problem in normal form.

the adopted plan in the absence of positive reasons for deviation (see Rabinowicz 1995).

3 Disadvantages of sophistication

The sophisticated approach is not without its problems. One difficulty, which would require a longer discussion, has to do with a controversial presupposition of the backward induction reasoning: The presupposition is that the agent confidently expects not only to act rationally in the future (which in itself is a strong assumption) but also to retain this confidence in her future rationality *come what may*. This is why she predicts, for example, that in the next-to-last choice node, she will expect to make the best move at the last choice node and thus will act accordingly. More generally, it is this expectation of continued confidence in her own rationality that allows her to make predictions about her future beliefs and consequently about her future choices, *even* at the nodes that can only be reached by irrational moves. Such predictions are needed for backward induction to go through. However, stubborn self-confidence, which ignores evidence about one's past irrational behavior, doesn't seem to be especially reasonable. It might therefore seem implausible for a rational agent to expect it.[1]

While this difficulty is serious, it doesn't arise as long as no branch of the decision tree contains more than two choice nodes – as in our example. There is then no later point at which the agent needs to retain her original confidence in her own rationality in order to make predictions about her future moves. Note that backward induction doesn't require such self-confidence at the terminal choice nodes.

A more worrying problem in the present context is that sophistication doesn't protect the agent against other forms of criticism. In particular, sophisticated agents who violate Independence will sometimes

(i) choose strictly dominated plans (McClennen 1990),

(ii) freely give up their freedom of choice, and

(iii) avoid costless information (Wakker 1988; Machina 1989).

Our example can be used to show this. Consider _Figure 35.1]] again. We have seen that the sophisticated agent will go down in the first move. But the difference between going down and up consists only in that, by going down, the agent gives up her power to choose X rather than Y if E is going to occur. And, in addition, she pays (ε) for having this freedom of choice taken away from her. Thus, (ii) holds. (i) holds as well,

[1] For a discussion, see Binmore (1987), Reny (1988, 1989), and Pettit and Sugden (1989). For a defense of backward induction against this objection, see Sobel (1993). For a restricted defense, limited to a particular sub-class of decision problems, see Rabinowicz (1998) and Broome and Rabinowicz (1999).

since plan t which is recommended by the sophisticated approach is strictly dominated by plan s: Whatever happens, whether or not E occurs and whether or not lottery Y would be won by the agent, she receives less (by ε) if she follows t rather than s. To show that (iii) holds as well, we would need to give the agent another alternative: to allow her, if she so prefers and at no extra cost, to make her move at the second choice node without knowing whether E has taken place. The sophisticated agent would opt for this, since it would make the attractive plan s performable. Thus, a sophisticated agent would willingly avoid costless information.

How is the sophisticated agent going to react to these observations? That (i) holds she will be prepared to accept with equanimity. It is true that s strictly dominates the plan t she adopts, but – she will point out – plan s is simply not performable. Dominance is of interest only if the dominating plan is not merely theoretically possible but also practically possible to execute. As far as (ii) and (iii) are concerned, she will argue that it is a prejudice to believe that information and freedom of choice can never be harmful – that we never have reason to avoid them.

How satisfactory is this reply? In a way, it is well-taken. In Homer's story, Ulysses had a good reason to give up his freedom of choice when he let his sailors to bind him to the mast. Likewise, his sailors had a good reason to avoid information when they plugged their ears in order not to hear the Sirens' song. But Ulysses and his crew acted as they did because they feared that the Sirens' song would distort their preferences. The case we discuss is different. The sophisticated agent who violates Independence does *not* expect her preferences to change upon learning that E has occurred. She will then prefer X (a safe large gain) to Y (a high chance of a larger gain), but so does she from the outset! She cannot offer the same excuse as Homeric heroes.

Thus, we are left with a lingering suspicion that something is not quite right with sophisticated choice after all.

4 Resolute choice

We might therefore want to consider another approach to dynamic decision making: *resolute choice*, advocated by McClennen (1990). Machina (1989) develops similar ideas.

As McClennen understands this approach, preferences of a resolute agent at later choice nodes are decisively influenced by the action plan she follows: She adjusts her preferences at later nodes to the adopted plan, so that at each later stage she prefers to follow the plan rather than to deviate. It is an "endogenous preference change" (McClennen 1990, section 12.7, and 2008, p. 132).[1] Thus, consider a resolute agent who prefers X to Y and YpZ to XpZ. If she confronts the dynamic decision problem we have described and adopts the attractive plan s, this causes her preferences change when she arrives to the second choice-node: While she initially prefers X to Y, she comes to prefer Y to X (and act accordingly) when she subsequently needs to choose between these two options.

Consequently, a resolute agent can afford to have preferences that violate Independence. In dynamic contexts, she is not going to be troubled by the difficulties that confront a sophisticated chooser. For her, all the theoretically possible plans are performable. She has therefore no reason to settle on dominated plans, to give up her freedom of choice, or to avoid information.

Unlike the sophisticated approach, the resolute approach accepts Reduction to Normal Form: The best plan in the normal form of a sequential choice problem is also best in the original problem in extensive form. Unlike the myopic agent, however, the resolute agent is not threatened by dynamic inconsistency. She not only adopts the best theoretically possible plan but also follows it through.[2]

Machina characterizes this approach somewhat differently. He emphasizes that a rational agent cannot ignore the history that has brought her to a given choice node. In her deliberation, she doesn't restrict her attention to that part of the decision tree that still lies in the future; she also considers how she has reached the point at which she is

[1] A more complex account of plan implementation, which opens up for counter-preferential choice, was suggested in McClennen (1997, pp. 238f). On that account, a resolute agent follows the adopted plan even if this conflicts with her preferences. Many economists who nowadays write about resolute choice interpret it very broadly as sticking to the plan, whether or not it means that one's subsequent preferences must undergo a process of adjustment (see Hey & Panaccione 2011; Hammond & Zank 2013).

[2] Empirical experiments suggest that resoluteness is by far the most common way of dealing with dynamic decision-making (Hey & Lotito 2009; Hey & Panaccione 2011). Though some of the subjects classified as resolute in these experiments might instead be wise choosers (see the next section).

now. Furthermore, she takes into account counterfactual developments – the branches in the tree that might have been actual had she or Nature made different moves in the past. This attention to history and to counterfactual histories might modify the agent's preferences and thereby cause her to stand by her original intentions.[3]

Both McClennen and Machina argue that the fundamental mistake of the sophisticated approach consists in its being purely forward-looking. What has been and what might have been doesn't count on this approach: Bygones are bygones.

Here's some notation that will be useful in what follows: If T is a decision tree and n is a node in that tree, let T_n be the *truncated* tree that is exactly like that part of T which starts at n. Note that T_n is a tree in its own right and not just a part of a larger tree.

According to McClennen (1990, sections 1.7 and 7.5) and Machina (1989, pp. 1622, 1639ff), there is an important assumption that underlies the sophisticated approach. Machina, following Hammond (1988), calls it Consequentialism, while McClennen refers to it as Separability. I will use McClennen's terminology. In Machina's version, Separability might be put as follows:

> **Separability** (Machina): If n is a choice node in a decision tree T, then the agent's preferences at n in T are the same as they would be at n in the truncated tree T_{n}.

At n in T_n, these preferences would be as they are at the root of T. As Machina clarifies, a separable approach

> consists of 'snipping' the decision tree at (that is, just before) the current choice node, throwing the rest of the tree away, and recalculating by applying the original preference ordering (or original preference function) to alternative possible continuations of the tree.
>
> (ibid., p. 1641f)

In McClennen's version (cf. his 1990, p. 122), Separability is a principle of rationality, or, to use his own term, a principle of "acceptability":

> **Separability** (McClennen): It is acceptable at n in T to continue on an action plan if and only if this continuation would form an acceptable plan in T_{n}.

[3] Not necessarily, though. Unlike McClennen, Machina does not stress the importance of the previously adopted action plans. What is particularly important for him is the need to attend to various *risks* the agent has borne in the past – even if those risks have never materialized. Attending to past risks would normally make one less prepared to engage in new risky gambles. In our example, however, this wouldn't help to make plan s performable. If the agent goes up and E occurs, she needs to attend to the past risk of receiving nothing (Z) had E not occurred. But this, if anything, should rather

Since what is acceptable, or rational, to do is determined by the agent's preferences, the two versions of Separability are closely connected.

It is Separability – pure forward-lookingness – that according to McClennen and Machina lies behind the sophisticated approach with its use of backward induction. When reasoning backwards, they suppose, one asks oneself each time what one would do, as a rational agent, in a truncated decision tree that starts at a node under consideration.[4] The pre-supposition is that this would tell the agent, given her expectation of her future rationality, what she would do in the original non-truncated tree were she to reach the node in question. It is because of Separability that attractive action plans, such as plan s in our example, get rejected by the sophisticated agent as not being performable. Such plans wouldn't be followed through if the agent is purely forward-looking.[5]

strengthen than weaken her preference for the safe X as compared with the risky Y.

[4] Harsanyi (1992, p. 369), who, unlike McClennen and Machina, embraces backward induction, agrees: "Subgame consistence" (his term for Separability) is an "essential component" of "backward-induction rationality." Subgame consistency states that the solution that in a larger extensive-form game G is prescribed for the subgame of G that begins at n coincides with the solution for the truncated game G_n.

[5] Sophisticated choice is not the only dynamically consistent decision policy on offer that satisfies Separability but violates Reduction to Normal Form (for agents who don't obey expected utility axioms). Ahmed (2016) has proposed such an alternative policy, which he calls *self-regulation*. But he only shows how self-regulation allows the agents with cyclic preferences to avoid dynamic inconsistency. It is unclear whether and how this approach can be extended to agents who violate Independence.

5 Wise choice

It seems to me that there is something reasonable in both sophisticated and resolute choice, but, at least as they have been described, they both seem too extreme. I want to plead for a conciliatory approach, which I call Wise Choice.

Wise choice is essentially sophisticated choice without Separability. Or, to put it the other way round, it is an approach that allows for resoluteness but doesn't require it. Thus, it doesn't assume Reduction to Normal Form.

Let me explain the main idea before giving a fuller characterization of this policy. As we have seen, in backward induction, I identify my best moves at later choice nodes relying on predictions about the preferences I expect to have at the nodes in question. Let n be such a node in my decision tree T. Now, it seems to me rather obvious that my preferences at n in T need *not* necessarily be the same as the preferences I would have had at n in the truncated tree T_n. My preferences at n in T might well be path-dependent: they might be influenced by what has happened prior to n and by what might have happened instead. Adherents of resolute choice are right on this point.

I can predict such changes in my preferences. Thus, I can show foresight and reason backwards without obeying Separability. In some cases, I might foresee that my future preferences will get adjusted to the previously chosen plan of action. Therefore, I might be able to predict that I will implement the plan – even though I wouldn't make some of the moves it prescribes if I didn't previously adopt it. In a truncated tree, my preferences (and thus also what I would do as a rational agent) might well be different.[1]

However, this rejection of Separability doesn't mean that I accept Reduction to Normal Form. The influence of the previously adopted plan on a wise agent's preferences at n need not always be strong enough to secure plan implementation. The best plan in the normal form might therefore still remain beyond the agent's reach even though Separability doesn't hold as a general constraint. In our example, the attractive plan s might not be performable for some wise agents: Were they to adopt this plan, they would prefer to deviate from it at the second choice node. In this approach, then, resoluteness is not required. Instead, it is best seen as a character feature that can vary in strength among different individuals: Some agents are more resolute than others and this variability is exhibited by wise agents as well.

But how does a wise agent predict her preferences at a future node n? We assume that she does not expect these preferences to be distorted. Therefore, she takes her future preferences at n to coincide with her current *conditional* preferences. More

[1] This means that many wise agents might outwardly behave like resolute choosers. There is thus

precisely, she takes them to coincide with her current preferences conditioned on the supposition that she would arrive at n. To illustrate, I can ask myself: Suppose I were to go up first, and E would then occur. I would then find myself in the second choice node. What do I now prefer with regard to that hypothetical situation? X or Y? If the answer is Y (even though I unconditionally prefer X to Y), I may predict that I would prefer Y if I were to reach that point. The assumption is thus that the agent's future preference in a hypothetical situation would be identical to her current conditional preference regarding that hypothetical situation.

How are conditional preferences to be understood, more exactly? Let A, B, and C be any propositions. The following definition of conditional preference seems plausible:

> A is preferred to B on the condition that C iff $A\&C$ is unconditionally preferred to $B\&C$.[2]

Just as A is more probable than B on the condition that C iff the unconditional probability of $A\&C$ is greater than that of $B\&C$.

There is, however, an important complication to consider in connection with the procedure I have just sketched. When the agent attempts to determine her conditional preference for n, that is, her preference on the hypothetical supposition that she will arrive at n, she must take into account how precisely she will reach that node. Thus, she must take into account the action-plan that will lead her to n. But the same choice node could sometimes be reached by following different plans, if they all prescribe the initial moves that lead to the node in question and start to diverge only at some later point. In our example, the agent might reach the second choice node by following either plan r or plan s. Her preferences at that node might well depend on which plan she has been following. Indeed, a choice node can also be reached by *deviating* from the adopted plan of action. In our example, the agent can form a plan to go down in the first node (plan t) but deviate from it and instead go up. Then, if E is going to occur, she will find herself in the second choice node as a result of this deviation from the plan she has adopted. How the agent arrives at a choice node may affect her preferences at the node in question. It is therefore necessary to include the originally chosen plan into the description of the hypothetical choice situation – to include it in the specification of the condition for the relevant conditional preference.

Consequently, we may characterize wise choice as follows:

First, one has to identify those theoretically possible plans that are performable, that is, that would be implemented if chosen. A plan is performable iff, if embarked upon, it prescribes best moves at all the subsequent choice nodes that can be reached by following the plan. Whether a move that belongs to the plan is best at its choice node

---

a need for caution regarding experimental findings that suggest the prevalence of resolute choice (see fn 10).

[2] To prefer one proposition to another means that one prefers the former being true to the latter being true.

is determined by backward induction based upon the agent's conditional preferences for this node and the succeeding choice nodes, where these preferences are *furthermore* conditioned on the hypothetical assumption that the agent has adopted the plan in question.[3]

In the second step, the wise agent chooses among performable plans one that leads to a best expected outcome. This choice is based on her original, unconditional preferences.

The present approach rejects Separability as a general condition on rational preferences, but it does not exclude that a wise agent's preferences may in fact be separable. If such an agent, in addition, violates Independence, she confronts the same problems as her sophisticated cousin: In some cases, she might opt for strictly dominated plans if the dominating plans aren't performable. She might also be willing to give up her freedom of choice and to avoid costless information. I don't think this makes her less wise.

Since resoluteness as a character feature might come in degrees, some wise agents will be resolute to a limited extent. For such agents, only those plans will be performable that do not require too radical preference adjustments at later choice nodes. Consequently, their optimal plans of action will sometimes differ both from the plans prescribed by sophisticated choice and from the plans prescribed by resolute choice.

To illustrate, consider a slight variation of the decision problem on which we have focused until now. In the new problem, if the agent goes up and E occurs, then at the second choice node, she has three alternative moves and not just two: she can opt for the safe large gain X, the slightly risky larger gain Y, *or* the significantly risky very large gain U. To fix the intuitions, suppose that X is a safe \$30,000, Y is a 98% chance of \$50,000, as before, while U is a 90% chance of \$100,000. Adding U as the third option at the second choice node means that an additional plan u becomes theoretically possible: to go up first and then, if E occurs, to opt for U. $\mathrm{P}(E) = p = 0.5$, as before.

Among the theoretically possible plans, let us assume that u is the most attractive: its expected outcome, UpZ, is preferred by the agent to the expected outcome of s (YpZ), which in turn is preferred to the expected outcome of r (XpZ). The increases in potential gains as one moves from XpZ to YpZ and from YpZ to UpZ amply compensate the relatively small increases in risk of ending up with nothing (this risk is 50%, 51%, and 55%, respectively). So, resolute choice would prescribe plan u. However, our agent is risk-averse: she prefers X to Y and Y to U. Sophisticated choice assumes Separability and thus prescribes plan t, as in our first example. Given Separability, only plans r and u are performable and of these two, the latter has a better expected outcome. Wise choice comes with a different recommendation if the agent's preferences

[3] Backward induction pre-supposes that one identifies best moves even for the future choice nodes that aren't reachable without deviation from the plan that is being considered. Even for such nodes, the best moves are determined on the supposition that a given plan has been adopted (but then given up).

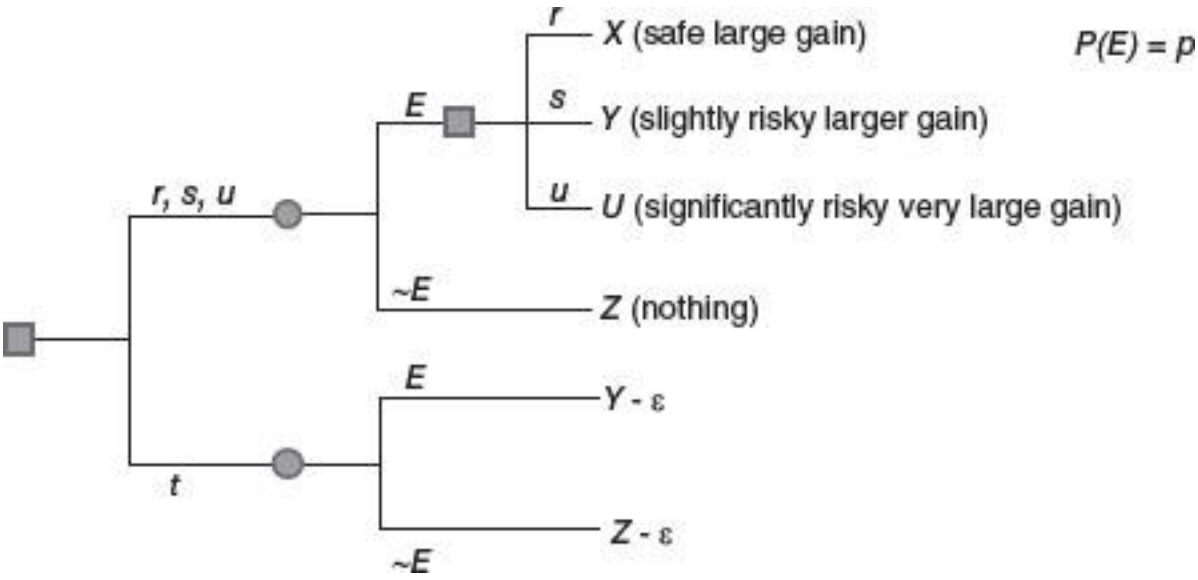


*Figure 35.2*Modified Allais problem in sequential form

aren't fully separable. Suppose she is to some extent influenced by whatever plans she adopts, but her resolve to follow the adopted plan is not absolute: even at later choice nodes, she still remains risk-averse, to some extent at least. In particular, were she to choose plan u, she would not stand by this resolution. The risk involved in opting for U is just too high, even on the hypothetical assumption that she would reach the second choice node while following plan u. So, u is not performable given the agent's limited degree of resolution. But this degree of resolution could still suffice to overcome her risk aversion if she instead were to adopt plan s, which requires her to choose Y at the second node. The reason is that the risk incurred by Y is minute. On the hypothetical assumption that she will reach the second node while following plan s, she prefers to stand by her plan and go for Y. Thus, plan s would be performable for this wise agent. Since s's expected outcome is better than that of the other two performable plans (r and t), this plan is optimal. In this case, then, wise choice recommends a plan that differs from those recommended by resolute and sophisticated choice (plans u and t, respectively).

Finally, it should be pointed out that wise choice can be generalized to cases in which the agent expects her future preferences to be *distorted* in some way, as in the case of an agent who expects to lose her restraint after a couple of drinks. Even with such expectations, she can still identify performable plans and choose among them on the basis of her current preferences. In determining whether a plan is performable, she would even in this case rely on predictions about her preferences at later choice nodes. The only difference is that the basis for such predictions will now have to be different from the one that is available for agents whose preferences change by conditionalization:

her future preferences won't be the same as her current conditional preferences with regard the future choice nodes.[4]

[4] A related issue, which I have not considered, is how the beliefs of an agent can change as she moves along a branch in the decision tree. This is relevant to the extent that her preferences over available moves at a given choice node are dependent on her beliefs at that node. For a rational agent, beliefs change by conditionalization on new information. But what if a wise agent can predict that her future beliefs won't change in this way? What if she can expect this kind of epistemic distortion? Well, then she must take this into consideration when she determines which of the plans at her disposal are performable. But the whole issue of beliefs in the context of backward induction is quite vexed. As I noted previously, it is necessary for backward induction that the agent retain her trust in her future rationality even at choice nodes she cannot reach by rational moves. At these nodes retaining trust in one's future rationality might in itself require violation of the conditionalization model for belief change.

6 But is wise choice really different from sophisticated choice?

Friends of sophisticated choice might raise an objection that wise choice is not an essentially novel approach.[1] Arguably, it is reducible to sophisticated choice if the decision problem that the agent confronts is appropriately re-described. This reduction can be achieved in four steps:

> (i) We should *add the planning node* at the root of the decision tree. At that initial node the agent makes a choice between plans. Each plan specifies how the agent is to act at subsequent choice nodes that are reachable by following the plan in question (possibly with Nature's help).
>
> (ii) We should *re-describe final outcomes* – make their descriptions more complete. For each final outcome, we should specify not only the direct payoff, such as, say, "safe large gain," but also everything else that matters about this outcome from the agent's point of view. If the history leading to this point matters to the agent, or if unrealized possibilities matter, they should be specified and included. If it matters whether the agent has followed her plan or deviated from it, this should be included.
>
> (iii) We should *apply backward induction to this expanded decision tree with re-described final outcomes*. Consequently, we now no longer need to rely on conditional preferences for future choice nodes: Since the final outcomes now are described in all relevant respects, the agent's preferences at future choice nodes can be assumed to coincide with her current *un*conditional preferences.[2] This means that Separability is going to be satisfied even for agents who care about history, unrealized possibilities and previously adopted plans.
>
> (iv) Finally, by moving backwards in this enriched decision tree, we arrive to the initial node in which a plan is to be chosen. A theoretically possible plan is *performable* iff its implementation would be supported by backward induction if the plan in question were chosen. A plan is *optimal* iff it is

[1] I am indebted for this objection to Kenny Easwaran and Jim Joyce.

[2] Thus, instead of comparing an outcome A with an outcome B on a condition C, one will now compare enriched outcomes $A\&C$ and $B\&C$. Given the suggested definition of conditional preference,

> performable and its expected outcome is most preferred in comparison with other performable plans. This is not quite the way we have defined optimal plans in our previous characterization of sophisticated choice, but the reason is that we previously thought of plans as branches of the decision tree rather than in the way we model them now – as alternative moves at the planning node.

Are plans that are not performable even available for choice in the planning node? One might well doubt it and argue that the agent cannot choose a plan without expecting that she will implement it if she makes this choice. And a sophisticated agent expects to implement a plan only if its implementation is supported by backward induction. On the other hand, to exclude unperformable plans from the decision problem, we would have to conduct backward induction before we even have identified the decision problem to which it can be applied. I think, therefore, that all theoretically listed plans should appear as moves in the planning node, even though the subsequent analysis of the decision problem can lead us to conclude that some of them are not really available for this particular agent. Thus, in our first example, not only plans r and t but also plan s should be listed in the planning node, even if, for the latter plan to be performable, the agent would have had to be more resolute than she actually is.

Note that, if some of the plans listed at the planning node are not performable, then an optimal plan need not be one choosing which would be recommended by backward induction at the node in question. For it might well be the case that backward induction recommends at that node choosing a plan that is not performable. Indeed, this choice might have excellent effects just because the chosen plan would not be implemented. To see this, consider the following variant of Kavka's Toxin Puzzle (1983). Suppose there is a toxin such that drinking it would have fatal effects. The agent knows that if she adopts the plan to drink the toxin, she will be generously rewarded by a large sum of money. The reward is offered *not* for the act of drinking but for the adoption of the plan to drink, regardless of whether this plan is going to be implemented. In the planning node, the agent has a choice between the plan to drink and the plan to abstain. At the second node, she will have a choice between drinking and abstaining. Backward induction recommends abstaining from drinking in the second node and adopting the plan to drink in the planning node. Choosing the plan to drink would have excellent effects, though only because this plan would never be implemented if chosen. (The agent has no use for the money if she is going to drink the toxin and die.) However, planning, and more generally intending, is incompatible with the belief that one won't do what one intends. Clearly then, we cannot apply backward induction in the planning node of the decision problem.[3]

---

these two comparisons are equivalent.

[3] Backward induction recommends adopting the plan to drink the toxin even in Kavka's original version of the Toxin Puzzle, in which the adverse effects of the toxin are mild and amply compensated by the reward for adopting the plan to drink. Also in Kavka's original version, adopting the plan to

Is the previously sketched reduction of wise choice to sophisticated choice satisfactory? It would seem so. It might, however, be objected that Separability is trivialized by such a maneuver. This condition was meant to express the future-directedness of sophisticated choice. But if historical features are allowed to be brought into outcome specifications, Separability no longer requires future-directness.[4] Also, the reduction implies that decision trees will have to be considerably more complex than in the standard presentations of decision problems. Thus, in our first example (Figure 35.1), in the planning node the agent first chooses between plans r, s, and t and only then confronts a subtree that looks like the tree in Figure 35.1. In this subtree, though, each final outcome includes not only the payoff but also, if it matters to the agent, the path through the tree that she has traveled, including the plan she has adopted, and the description of paths she might have taken instead. Still, if we are prepared to accept this greatly increased complexity, both in the number of branches and in the outcome specifications, and if we have nothing against Separability being trivialized, the reduction seems to work.

But there is a limit to this reduction. As pointed out at the end of the preceding section, wise choice admits of generalization to cases in which the agent expects her future preferences to be distorted, as seen from her current point of view. Sophisticated choice leaves no room for such generalization. It determines best moves at all later choice nodes relying on the agent's current preferences. So, at this point, wise choice leaves sophisticated choice behind.

It is time to sum up. In this chapter, I have considered potential dynamic inconsistency that threatens agents who violate expected utility axioms, in particular the axiom of Independence. I have presented three policies that deal with this threat: sophisticated choice, which assumes Separability; resolute choice, which assumes Reduction to Normal Form; and wise choice, which assumes neither. The latter is the policy I favor. Like sophisticated choice, wise choice makes use of backward induction,[5] but in this procedure relies on the agent's conditional, rather than unconditional, preferences. Thereby, the potential influence of the past and of the plans one has adopted can be taken into account in practical deliberation. Depending on their conditional preferences, wise choosers might sometimes behave like sophisticated choosers, sometimes like resolute choosers, and sometimes like neither. I have finally considered whether and to what extent wise choice can be seen as a version of sophisticated choice. For

drink is not performable. The agent knows that she doesn't care about being true to her resolutions, and thus that she won't drink the toxin – whatever plan she now were to adopt. His version leaves it open, however, that the plan to drink the toxin would be performable for a more resolute agent, who cares more about her commitments. In my version, the cost of such steadfastness is just too high.

[4] I am indebted to Peter Vallentyne for pressing this point.

[5] Although see Peterson and Vallentyne (2018), where it is suggested that reliance on backward induction is not essential for either sophisticated or wise choice. Backward induction involves making definite predictions about one's future behavior. Peterson and Vallentyne suggest a generalization in which such specific predictions are replaced by probability assignments to one's future moves.

such a reduction, a re-description of the decision problem is required in order to recover Separability without losing the advantages of the wise choice approach. There are costs and limits to this reductive enterprise.

Acknowledgments

This chapter has been long in the making. Its earlier versions were presented at a workshop on self-prediction in Cambridge in May 2015, at the Institute of Futures' Studies in Stockholm in February 2016, at a seminar at Lund University in April 2017, and at the Round Table on Dynamic Consistency Beyond Expected Utility that was part of the conference on Foundations of Utility and Risk (FUR) in York in June 2018. I am indebted to the participants of these events for useful suggestions. Special thanks are due to Kenny Easwaran and Jim Joyce, who have made me re-think my proposal and add the final section. I have also received very helpful comments from Arif Ahmed, David Alm, Lara Buchak, Robin Cubitt, Peter Hammond, Johan Gustafsson, Martin Peterson, Peter Vallentyne, and an editor of this volume, Kurt Sylvan.

References

Ahmed, Arif (2016), "Exploiting Cyclic Preference", *Mind* 126: 975–1022.

Ainslie, George (1992), *Picoeconomics*, Cambridge: Cambridge University Press.

Allais, Maurice (1953), "Le comportement de l'homme rationnel devant le risque: Critique des postulates et axiomes de l'école Américaine", *Econometrica* 21: 503–546.

Binmore, Ken (1987), "Modelling rational players: Part 1", *Economics and Philosophy* 3: 179–213.

Bratman, Michael (2012), "Time, rationality, and self-governance", *Philosophical Issues* 22: 73–88.

——— (2014), "Temptation and the agent's standpoint", *Inquiry* 57: 293–310.

Broome, John, and Wlodek Rabinowicz (1999), "Backwards induction in the centipede game", *Analysis* 59: 237–242.

Hammond, Peter J. (1988), "Consequentialist foundations for expected utility", *Theory and Decision* 25: 25–78.

Hammond, Peter J., and Horst Zank (2013), "Rationality and dynamic consistency under risk and uncertainty", in M.J. Machina and W.K. Viscusi (eds.), *Handbook of the Economics of Risk and Uncertainty*, vol. 1, Amsterdam: Elsevier: 41–97.

Harsanyi, John C. (1992), "Game solutions and the normal form", in C. Bicchieri and M.L. Dalla Chiara (eds.), *Knowledge, Belief, and Strategic Interaction*, Cambridge: Cambridge University Press: 355–376.

Hey, John D., and Gianna Lotito (2009), "Naive, resolute or sophisticated? A study of dynamic decision making", *Journal of Risk and Uncertainty* 38: 1–25.

Hey, John D., and Luca Panaccione (2011), "Dynamic decision making: What do people do?" *Journal of Risk and Uncertainty* 42: 85–123.

Kavka, Gregory S. (1983), "The toxin puzzle", *Analysis* 43: 33–36.

Machina, Mark J. (1989), "Dynamic consistency and non-expected utility models of choice under uncertainty", *Journal of Economic Literature* 27: 1622–1668.

McClennen, Edward (1990), *Rationality and Dynamic Choice*, Cambridge: Cambridge University Press.

——— (1997), "Pragmatic rationality and rules", *Philosophy and Public Affairs* 6: 317–344.

——— (2008), "Exploitable preference change", in T. Grüne-Yanoff and S.O. Hansson (eds.), *Preference Change*, Dordrecht: Springer: 123–138.

Paul, Sara K. (2014), "Diachronic incontinence is a problem in moral philosophy", *Inquiry* 57: 337–355.

Peterson, M., and P. Vallentyne (2018), "Self-Prediction and Self-Control", in J. Bermudez (ed.), *Self-Control, Decision Theory, and Rationality*, Cambridge; Cambridge University Press: 48–71.

Pettit, Philip, and Robert Sugden (1989), "The backward induction paradox", *The Journal of Philosophy* 86: 169–182.

Rabinowicz, Wlodek (1995), "To have one's cake and eat it, too: Sequential choices and expected-utility violations", *The Journal of Philosophy* 92: 586–620.

——— (1997), "Wise choice: On dynamic decision-making without independence", in E. Ejerhed and S. Lindström (eds.), *Logic, Action, and Cognition*, Dordrecht: Kluwer: 97–112.

——— (1998) "Grappling with the centipede. Defence of backward induction in BI-terminating games", *Economics and Philosophy* 14: 95–126.

——— (2000), "Money pump with foresight", in M.J. Almeida (ed.), *Imperceptible Harms and Benefits*, Dordrecht: Kluwer: 123–154.

——— (2001), "A centipede for intransitive preferrers", *Studia Logica* 67: 167–178.

——— (2008), "Pragmatic arguments for rationality constraints", in M.-C. Galavotti, R. Scazzieri and P. Suppes (eds.), *Reasoning, Rationality and Probability*, Stanford: CSLI Publications/The University of Chicago Press: 139–163.

Reny, Philip J. (1988), "Rationality, common knowledge and the theory of games", Ph.D. diss., Princeton University, Princeton.

——— (1989), "Common knowledge and games with perfect information", in *PSA 1988: Proceedings of the Biennial Meeting of the. Philosophy of Science Association*, vol. 2, East Lansing, MI: Michigan State University: 363–369.

Skyrms, B. (1993), "A mistake in dynamic coherence arguments?" *Philosophy of Science* 60: 320–328.

Sobel, Howard (1993), "Backward induction arguments in finitely iterated prisoners' dilemmas: A paradox regained", *Philosophy of Science* 60: 114–133.

Wakker, Peter (1988), "Non-expected utility as aversion to information", *Journal of Behavioral Decision Making* 1: 169–175.

36. The norms of practical reasoning

Jennifer M. Morton and Sarah K. Paul

It is possible to act unthinkingly, but many of our actions are the result of reflection about what to do. Let us call this process of thinking about what to do, concluding either in an intention or an action, 'practical reasoning' or 'practical deliberation'. We generally do not need to deliberate about whether or how to brush our teeth before going to bed, but we do deliberate about questions like whether to have a child or how to get accepted to law school. And in addition to evaluating actions for their merits and persons for their characters, episodes of reasoning can be evaluated as more or less good. The central question for this chapter is: in virtue of what is it the case that a person deliberated well or poorly? How ought we to reason about what to do?

With a dash of oversimplification, we can usefully divide the literature on this question into a rough dichotomy. One dominant view is that there are internal standards of good practical reasoning as such, according to which an agent's reasoning can be evaluated independently of whether it led to performing the correct action. This thought is standardly fleshed out in terms of formal rational requirements and permissions that liken practical deliberation as closely as possible to deductive logic. For instance, many have thought that there is a rational requirement enjoining agents to intend the means believed necessary to their intended ends. This is a formal principle, in the sense that it applies independently of what the end is and what the means turn out to be. But it is also formal in the sense that it applies to all rational creatures engaged in the activity of practical reasoning (or at least to all finite creatures who must bother taking means to our ends), regardless of any other property of those creatures.

The other prevalent approach holds that good deliberation is nothing over and above responding correctly to the reasons we have to do or intend particular things. Practical reason is understood simply as the capacity to respond appropriately to value, or to what we have most reason to do, and what matters is only that it leads us to do so. Proponents of this view are skeptical that there are internal standards of practical reasoning that play any direct role in determining or explaining what an agent should think or do. We can use formal rules as heuristics if we like to evaluate each other's deliberation, but they are not genuinely normative – the agent herself is free not to care about conforming to them. In this way, they are akin to the rules of etiquette.

In what follows, we survey these two approaches and argue for a third, "ecological" alternative with roots in the tradition of Herbert Simon's "bounded rationality" theory (1955). There are norms of practical deliberation that have distinctive authority over us, in that we ought to be guided by them in reasoning irrespective of the conclusion they lead to in any particular case. We are habit driven, cognitively limited agents who will do better to follow general rules. However, the status of these rules as norms is not independent of the good effects of accepting and being guided by them; whether a given agent should be disposed to reason using a particular norm depends on the general advantage of doing so. In our view, this is an empirical question and will be relative not only to a given agent's cognitive capacities but also to the context she normally operates in. The ecological approach avoids the implausible claim that the authority

of practical norms is independent of whether they produce good results, while also avoiding the collapse into supposing that good deliberation is nothing over and above reasons-responsiveness.

I Norms as categorical rational requirements

According to the first approach, good practical reasoning consists in non-accidentally coming to satisfy certain formal requirements. Abstracting away from the details of any particular theory, one plausible example of such a requirement is the Instrumental Principle mentioned earlier: 'will (or intend) the means believed necessary for your ends'. Another example is the Enkratic Principle: 'if you believe you ought to F, intend to F'. A third candidate is the Consistency requirement: 'if you intend to F, do not intend anything you believe to be incompatible with F-ing'. Fourth, some think that there is a requirement enjoining Stability in one's intentions over time, absent some event (forgetting, or gaining significant new information) that releases the agent from this demand: 'if you intend to F at some future time, and no "canceling event" occurs, continue to intend to F.' Finally, those working in the Kantian tradition will add the Categorical Imperative to the list: 'act only in accordance with that maxim through which you can at the same time will that it become a universal law'.[1]

These formulations are only rough approximations, and there is much disagreement over precisely how they should be articulated. For instance, these examples are formulated as having "narrow scope," which means that whenever one satisfies the antecedent of the relevant conditional, one is required to satisfy the consequent. However, John Broome (1999, 2007a) and others (e.g. Way 2011) have argued that they should instead be understood as taking "wide scope" over the whole conditional or combination of attitudes, which prevents the detachment of any particular conclusion that is required. In other words, a wide-scope version of the Instrumental Principle forbids the combination of "intending E, believing that M must occur if E is to occur, believing that M will not occur unless one intends M, and lacking the intention to M." This requirement can be satisfied in multiple ways, which means that no particular intention, belief, or lack thereof is required. But whether wide-scope or narrow, the general thought behind the formal approach is that reasoning is correct just in case it brings

[1] Some views will further distinguish between "process requirements" and "state requirements," where state requirements instruct us to be a certain way, while process requirements instruct us to do something (Kolodny 2007). For our purposes, this distinction will not matter. We are interested in views that take some such principles to be requirements on reasoning, either because they are process requirements that direct us to reason in a certain way or because good reasoning consists in coming to satisfy the state requirements.

one to satisfy the rational requirements (and perhaps there are additional constraints on the permissible ways of doing so). The reasoning can be evaluated independently of whether the reasons or values at stake adequately support the action or intention that is the conclusion.

Since these standards of correct reasoning are not derived from any substantive result they might lead to, the question arises of where they come from. One school of thought is what we might call Intuitionism, in which no principled account can be given of why the requirements are what they are. Broome follows Thomas Nagel in arguing that

> One cannot discover or justify the principles which specify those requirements by deriving them from the concept of rationality, since it is precisely those requirements which define the concept, and they must be rendered plausible as requirements independently.
>
> (2013, 150)

He further claims that he can see no way to do this by appeal to the nature of the mental states they are concerned with or with some further purpose that rationality has. Rather, we must go largely on our intuitions.

The problem for the Intuitionist is that once the rational requirements are spelled out by appeal to intuition, there is an open question of why we should care about them. Does the fact that rationality requires one to F give one any *reason* to F? It is easy to think of cases in which conforming to the rational requirements leads the thinker to a conclusion that she has no moral or prudential reason to draw. For instance, a deeply mistaken belief about what she ought to do will only be exacerbated by the Enkratic principle, which enjoins her to actually intend that thing. Merely pointing out that her *reasoning* was correct in such cases should be cold comfort, akin to pointing out that she displayed good etiquette while doing something colossally evil or stupid. The Intuitionist can of course simply assert that the rational requirements have non-derivative normative authority over us without offering a further explanation, but this is unsatisfying (and Broome does not endorse this kind of "dogmatic rationalism," leaving open whether rationality is normative) (Broome 2007b).

A second, "Cognitivist" approach passes the buck over to theoretical rationality. The aim of such a strategy is to trade two normative problems for one. This kind of view depends on a controversial claim about the nature of intention: either that it is a belief or that it is a complex attitude involving belief. Gilbert Harman (1976) proposes that if intending to F entails believing that one will F, then we can reduce the practical requirement that one's intentions be mutually consistent to the theoretical requirement that one have mutually consistent beliefs (see also Velleman 2007). He further suggests that the Instrumental Principle might be explained in terms of a theoretical requirement prohibiting explanatory gaps in one's beliefs: one ought not to believe that one will do E while lacking a belief about how E will happen. If E will

only happen if the agent takes means M, then he must form the belief that he will M, thereby intending to M. Setiya (2007) defends a version of this strategy that appeals to the Closure principle on belief ('You should [if you believe that E and that if E, M, believe M'), while Wallace (2001) argues that the strategy can work even if intentions merely entail the belief that the intended end is *possible*.

The strategy of grounding the practical requirements on intention in corresponding theoretical requirements on belief is only as plausible as the claim that having an end entails believing that one will achieve that end. There are good reasons to think this is false, however. Many of our most important ends, like becoming a successful novelist or marrying for life, are very difficult to achieve, and it is possible to have those ends while maintaining doubt that one will succeed. More mundanely, it is possible to forget even to try to carry out an intention one has (such as to stop and get gas on the way home), and this is something we can anticipate about ourselves (Bratman 1987). A second problem for this strategy arises from the need to distinguish between means and foreseen side effects. If your end is to get to Auckland and the necessary means is to fly on an airplane, this will release pollutants into the atmosphere. We want the result that the requirements of practical rationality apply to the means of flying and not the side effect of polluting, but both are things that you believe must happen if you are to arrive in Auckland (Paul 2011). Finally, the question of why we should care about these requirements can simply rearise in the context of theoretical rationality. What is to be said in favor of being coherent in our beliefs, in addition to responding correctly to our evidence?

A third school of thought argues that the rational norms are constitutive of something else we are committed to. For instance, Christine Korsgaard (2008) argues that they are constitutive of being an agent: a unified, autonomous being that can make something happen. The Categorical Imperative articulates what it is for the will to be autonomous, and the Hypothetical Imperative (Instrumental Principle) articulates what it is for the will to be efficacious. On this kind of neo-Kantian view, it makes no sense to ask whether willing in accordance with these norms will always produce a result that is independently supported by what we have reason to do. Rather, an action is correct *only because* it has the form required by the two Imperatives. To try to question their normative authority would be to question whether to be an agent with an autonomous will, and this is not something we can do in a serious way.

This Constitutivist approach fares better than the Intuitionist or the Cognitivist in accounting for the normativity of rational requirements, but it faces its own difficulties. A major worry is that these formal requirements are simply too thin to place significant constraints on how we can correctly deliberate or to guide us to any substantive conclusions. With a bit of ingenuity, it seems possible to craft a maxim that satisfies the two imperatives for just about any action. This is Hegel's 'empty formalism' objection (Wood 1990). A possible response to this worry is to appeal to a rich conception of agency that leaves substantial room between minimally counting as an agent and exemplifying agential excellence. This might allow the view to derive further, more con-

tentful norms that are constitutive not just of being an agent but of being an excellent agent (Velleman 2009). However, this risks the status of those norms as categorical (for all finite rational creatures), since it will be controversial that all minimally rational beings are committed to being agents according to the rich conception being offered (Tiffany 2012).

There are further possible variations on these attempts to understand good practical reasoning in terms of conforming to categorical rational requirements, but we suggest that all such attempts will face a form of the dilemma exhibited in this section. Either the view in question allows that there is some independent normative standard by which we can evaluate the conclusion of reasoning, or there is not. If there is not, and the norms of reasoning solely determine whether the conclusion is correct, the problem is that any requirements that are universally applicable are also too thin to do the work of sufficiently constraining the output. If there is, then there will be cases in which adhering to the requirements will lead us to act in a way that is incorrect according to the relevant standard (act against our reasons, say), raising the question of why they should have any distinctive normative authority over us. We examine the latter objection in more detail in the next section.

II Skepticism about the normativity of practical reasoning

The difficulty of specifying rational requirements that are immune to examples in which one would do better to violate them has led some philosophers to conclude that conforming to rules like Consistency and the Instrumental Principle has no distinctive normative significance. An episode of deliberation is good, on this view, only insofar as it leads to the right results: intending or doing that which we have most or sufficient reason to do. In other words, deliberation is merely a tool for responding correctly to reasons.

For instance, Joseph Raz (2005) argues that the Instrumental Principle is a "myth," in that there is no reason as such to conform to it. In his view, "with creatures capable of reasoning about ends, reasoning about means is not distinctive and special, but part and parcel of our general rational functioning" (28). Agents who respond correctly to reasons or value might end up with attitudes that are in conformity with the Instrumental Principle, but this would be a side effect of responding correctly to reasons rather than an achievement specific to good reasoning. There is a principle in the neighborhood, the Facilitative Principle, that directs us to facilitate that which has value. But this principle has nothing to do with the attitudes an agent starts with or with deliberation as such. Rather, it is a principle describing how reasons for valuable ends transmit to reasons for taking the means to those ends.

Similarly, Niko Kolodny (2008) argues that the Consistency requirement on intention has no distinctive normative significance for us. He points out that a disposition toward consistency does not itself make us any more likely to adopt the attitudes that we have reason to have. Arbitrarily resolving a conflict between incompatible intentions is just as likely to lead one to abandon the wrong intention as to abandon the right one – and maybe it is the belief that the intentions are incompatible that is erroneous. Since we should care only about having the attitudes best supported by our reasons, and the rational requirements do not bring us closer to that goal, we have no reason to be directly concerned with satisfying them. Kolodny (2005) concedes that being in conflict with a requirement can alert us to the possibility that a mistake has been made, but only the reasons themselves should guide us in resolving it.

If the Instrumental and Consistency Principles fail to guide us toward the correct action in some cases, we might think the Enkratic Principle would fare better. Agents who act against their own best judgment are akratic, and akrasia has long been thought the paradigm case of criticizable irrationality. But even this norm has been subject to a

powerful critique by Nomy Arpaly. She points out that when an agent's beliefs about what she ought to do are mistaken, flawless deliberation in conformance with the Enkratic Principle will leave her worse off than if she had not deliberated enkratically at all (2000, 2003). The classic example is Huckleberry Finn, who is convinced that he ought to turn Jim in as a runaway slave but akratically fails to do so. Arpaly argues that by failing to satisfy the Enkratic norm, Finn ends up doing what he in fact has most reason to do and thus brings about a far better state of affairs than if he were to enkratically betray Jim.

Finally, this form of critique has been used to question the normative authority of deliberation as such. Arpaly and Timothy Schroeder (2012, 2013) argue that deliberation has been overvalued as the paradigm of acting for reasons and that we would in many cases have done better not to deliberate at all. Like Raz and Kolodny, they take agency to be the capacity for reasons-responsiveness and argue that deliberation is a useful but imperfect tool that assists us in this capacity. It aids us with overcoming cognitive limitations like accessing stored information, chaining together inferences, and overcoming distractions but is not in itself a source of reasons or essential to our capacity to act on them. They write:

> We can think of no realistic way in which human beings could have achieved the insights in philosophy, the arts, and the sciences that we have achieved without deliberation. So we celebrate deliberation. But we celebrate it as a powerful tool well suited to ameliorating our human weaknesses as imperfect ND [non-deliberative] reason-responding agents. We are unfortunate in needing deliberation but fortunate in having the tool that we need. Deliberation is a beautiful tool for its purpose. Philosophers go wrong only when they misconstrue deliberation as, not a tool in, but the foundation of, our power to think and act for reasons.
>
> (2012, 238)

These skeptics about the normative authority of deliberative activity and its associated principles are motivated by a common thought: that there are no rules that any rational creature can follow without sometimes being led astray (unless the principle is so toothless that it does no work in guiding us). So far, we are sympathetic to this objection. However, it does not follow that deliberation must turn out to be a mere crutch or that there are no deliberative norms that have authority independently of any particular result they produce. Rather, the solution is to embrace the contingency of any such norms and to evaluate them in a global, context-dependent way rather than on a case-by-case basis. We explain in the next section.

III Contingency and context

The view we favor has its roots in the bounded-rationality tradition of Herbert Simon (1955). Against the skeptics in Section II, we think norm-guided practical deliberation is more than just a crutch or heuristic. But against the categorical view articulated in Section I, we deny that the normative significance of deliberative rules derives from their status as purely structural requirements modeled on formal logic. The norms we ought to be guided by must be sensitive to particular features of our agency and our circumstances, and they cannot ensure that we will get the right result in any particular case. Still, they are indispensable to our form of agency.

First, a side note. As mentioned in Section I, there is a deep metaethical divide between those who think that reasons for action exist independently of the perspective of practical reason and those who do not. The skeptics in Section II fall squarely on the first side, and the alternative we articulate in this section is addressed primarily to them. However, it is worth mentioning that many of the features of the view we will sketch could be accepted by those who fall on the other side of the divide. We will return to this point at the end of this section. At this stage of the dialectic, we will assume for the sake of argument that reasons for action do exist independently of good practical reasoning, whether they are grounded in our desires and ends or in mind-independent facts about value. We deny that that the "norms" of deliberation must fail to be genuinely normative on this kind of view or that the value of deliberation must collapse into its (highly imperfect) usefulness in producing the right result on particular occasions.

Imagine a creature who is non-deliberatively reasons-responsive. It is sensitive to the reasons it has in a quasi-perceptual way, without asking itself the reflective question 'what to do?' When it witnesses an injustice, it intervenes. When it encounters a Rothko, it admires the beautiful, vibrant colors. When it is hungry, it eats. Arpaly and Schroeder argue that if such a creature's capacity for non-deliberative reasons-responsiveness were flawless, it would have no use for deliberation. They conclude that because the value of deliberation is merely contingent, "reason is not what makes it possible to think or act for reasons" (2013, 52). However, we should ask why the role of deliberation in the lives of flawless reasons-responders entails anything at all about the role it plays for flawed creatures like ourselves. We face a number of problems that such creatures lack, and some of these problems are not merely diminished forms of capacities they possess. It might be that for us, deliberation not only enhances our capacities but actually enables a certain kind of reasons-responsiveness that we would not otherwise have access to.

For instance, suppose that our non-deliberative reasons-responder has no issues with working memory, attention, or computational complexity but is subject to a fluctuation in his preferences and evaluative judgments over time due to temptation. In the morning, he sees that he has most reason to knock off work at 8pm, watch two hours of TV, and go to bed by 10pm. But that night, temptation causes his preferences and evaluative judgments to reverse themselves, whereupon he sees that he has most reason to work until 10pm and watch TV until 3am. The next morning, of course, he deeply regrets this behavior. The problem is that his capacity to respond directly to the reasons as he sees them will not help him keep to a better schedule, since that is already what he is doing at each point (Holton 2009). Perhaps this is already enough to show that he is not a flawless reasons-responder, since such a creature would be sensitive even to cross-temporal reasons that conflict with its present evaluative perspective. But if so, such creatures are very different from us indeed and possess a faculty for detecting reasons that is mysterious at best.

For creatures like us, it is only from a perspective in which we step back from our present desires and evaluative judgments and reflect on the pattern that we would prefer to exhibit over time that the problem even becomes visible. But this just is the perspective of reflective deliberation. From this perspective, our intemperate agent can adopt a policy or resolution aimed at settling for himself that he will quit work at eight and go to bed at ten, regardless of what he prefers or judges he ought to do at those times (Bratman 1996). Such a policy or resolution does not work by acting as a further consideration bearing on whether to work, watch TV, or sleep. Rather, the agent must see it as having the authority to settle the matter for him in advance without any further exercise of his non-deliberative capacities to respond directly to reasons. For this, a deliberative norm enjoining stability in one's policies and resolutions over time is precisely what is needed. The agent can see his policy as answering the question of what schedule to keep because he deliberatively settled on it and because he accepts a norm according to which this suffices to close the question (again, absent some canceling event like significant new information). The possibility of acting unreasonably by following such a policy does not show that we should *never* be guided by them, since we would then never succeed in overcoming diachronic temptation (see also Paul 2014).

A similar point applies to creatures facing choice situations that are normatively underdetermined. We are such creatures; we often have multiple options available to us that are incommensurably valuable or "on a par" (Chang 2002). The capacity for reasons-responsiveness alone will not point us to a unique course of action in such cases. The skeptic might grant that in these cases, we can simply pick one of the options in a non-deliberative way. But the problem is to explain how picking could itself have any normative significance: why follow through with the option you have picked, rather than switching to some other option that is just as valuable? In other words, how is commitment possible in the face of multiplicity of value (Raz 1999; Bratman 1996; Chang 2009)?

The skeptic tends to try to address this problem by locating some additional reasons in the neighborhood. Once we have intended some course of action, we have often raised expectations in others about what we will do and begun to invest resources in that plan (Scanlon 2004; Kolodny 2011). These facts can make switching to another option more costly, changing the balance of the relevant reasons. But frequently, the cost will not really be enough to render the other option(s) outright unreasonable to pursue; that an agent's friends and family will be surprised that she is dropping out of a Philosophy Ph.D. program to attend law school, and that she will be out the application fee she paid for the Ph.D. program, do not suffice to rule out law school as a good option. Moreover, sometimes the goals we settle on are challenging activities that require a long-term investment, such as writing a novel or running a marathon. As we encounter the frustrating setbacks that are characteristic of pursuing difficult goals, the alternatives we previously set aside will start to seem better and better. Thus, the concern is not really that we will switch mindlessly between two valuable activities or relationships for no reason. We will be inclined to switch because we begin to think that a previously discarded option would be an easier path to an end that is also good. But if we too easily abandon our goals in light of temporary setbacks, we will have no access to the distinctive value of very difficult achievements and long-term relationships (Morton and Paul 2019).

What can give us access to this kind of value is the disposition to reason with norms that favor commitment or "grit." The agent must see the fact that she considered the various options and committed to one of them as having some additional normative significance for her. Again, this might consist partly in a norm of stability that enjoins resistance to changing one's mind in the absence of significant new reasons. It might also involve epistemic norms that are selected to favor perseverance in the face of difficulty by permitting some degree of epistemic resilience in the face of evidence that one might not succeed (Morton and Paul 2019). Deliberation itself plays an important role because it is crucial that the agent see herself as having committed to one option *as opposed to* the other alternatives that she actively considered. Otherwise, being inclined after a while to switch to one of the alternatives will be akin to getting new information about one's options rather than something that has already been ruled out by the commitment she made.

We take the existence of temptation and normative underdetermination to show that deliberation in accordance with certain norms is not only an enhancement of our existing capacities for reasons-responsiveness but essential for our access to certain kinds of diachronic value. Even the otherwise flawless reasons-responder needs deliberation to solve these problems. However, the claim is not that following such norms will never lead us to the wrong result or that there could not be cases in which the justification for the norm does not apply. We grant the skeptic that counterexamples are always possible to devise. What, then, of the skeptical conclusion that these deliberative rules fail to be genuinely normative for us? Here, we suggest that the skeptic has a mistaken view of what it takes for a deliberative disposition to be justified as a

norm. A successful response to the 'why be rational?' question need not consist in an argument showing that following the norm invariably leads to the right result or that there is at least some reason to conform in each case to which the norm applies. It is enough to justify it at the global level: to show that we will be better off in general in virtue of being guided by that norm when reasoning as opposed to all other candidates or no such norm at all. And it is consistent with such a justification that sometimes, we ought not to be guided by these norms because we ought not to deliberate at all.

This is a version of the "two-tier" strategy articulated by Michael Bratman (1987) (though he himself no longer accepts its adequacy) and has affinities with Peter Railton's "sophisticated consequentialism" (1984). At the heart of this view is a psychological claim about how deliberative norms work. If the role of these norms is to take us from premises to conclusions, they cannot function as additional premises in reasoning about what to do, as Lewis Carroll's "Achilles and the Tortoise" shows (1895). Rather, they consist in *habits* or *dispositions* to reason in a certain way. They operate in the background of deliberation, giving it its structure. Although we can articulate them to ourselves and assess them, they cannot be the object of reflective scrutiny on a case-by-case basis, on pain of paralyzing our ability to deliberate altogether. Not only must deliberation be suspended whenever the rules of deliberation are called into question, but our cognitive and temporal limitations would make the attempt to continually tinker with them far too costly.

Therefore, when we ask questions like "Why be consistent?" we are really asking "Why be in the generally unreflective habit of being consistent?" According to the two-tier model, a positive answer to this question will suffice to justify being deliberatively consistent in any particular case, *even if* it is a case in which there is nothing otherwise to be said for consistency. Critics have objected that this amounts to a form of "superstitious rule worship" (Smart 1956). We think this objection lacks the bite that many have attributed to it, at least when applied to reasoning.[1] While it is true that there are possible circumstances in which we would do better to have incompatible intentions, fail to take the means to our ends, ignore our resolutions, and give up on the goals we have committed to, these circumstances will be relatively infrequent and abnormal. It will be no simple matter to be sure that one is in such circumstances, especially given the distorting effects of temptation, sour grapes, and despair. It requires an investment of time and energy to try to identify the exceptions, and this is a cost

[1] The objection was originally leveled at Rule Utilitarianism. Whether it succeeds in that context, the project of justifying deliberative norms is different than the project of explaining what makes it the case that an action is morally correct. For the latter project, the Utilitarian accepts a standard of moral goodness – "maximize happiness for the greatest number," say – and then attempts to justify the acceptance of certain rules in light of that standard. When the prescription of the rule deviates from the standard, all the justificatory force plausibly lies with the standard and not the rule. But in the case of deliberation, the norms are not being justified entirely with an eye to their consequences. The justification comes in part from how well they serve the function of deliberation in creatures who have need of it. Therefore, we think the "rule worship" objection does not translate straightforwardly between the two paradigms.

that the benefits will often fail to outweigh. Further, most of us tend to be over-rather than under-confident in our powers of discernment, and will be liable to think we are in exceptional circumstances when we are not.

Where critics see rule worship, therefore, we see epistemic humility: better to err on the side of being disposed simply to stick to the norms that have been justified by the general advantages they confer rather than to be on the lookout for exceptional circumstances. A good analogy is the enterprise of investing in the stock market. Even though some stocks really do outperform the market as a whole, most investors will do far better by investing in an index fund rather than trying to pick those winning stocks, since it is extremely difficult to do. This is well known, and yet many investors continue to try to pick their own stocks or pay hefty fees to equally fallible hedge-fund managers to do it for them. This overconfidence in their ability to identify the exceptional stocks almost always leads to a significantly worse performance over the long run.

Similarly, we limited and fallible human agents will almost always do better to deliberate habitually in accordance with the norms we have general reason to accept. Of course, it might be possible to violate the norms while believing that it is an 'exception' case and get the right result. Must we then say that the agent is subject to criticism for this? It depends on whether the agent genuinely *knew* he was in an exception case or simply got lucky. We have suggested that such knowledge will be difficult to get, since the agent's belief will often fail to be safe – he could easily have been in a qualitatively similar situation in which he would have done better to conform to the norms. If he does not know, then it is appropriate to criticize him for reasoning badly even if we grant that he acted rightly as a result. If he does know, then he has reasoned permissibly, since his knowledge undercuts the normative force of the rule(s) he violated.

Finally, more needs to be said about how a candidate norm should be evaluated at the level of a disposition to deliberate in a certain way. Again, at this point in the dialectic, we are assuming for the sake of argument that there are reasons for action that exist independently of the activity of practical reasoning. We have argued that deliberation is not only a valuable tool for responding to these reasons; in some situations, it gives us access to reasons that we otherwise would not have had. Furthermore, the norms that a given agent should deliberate with should be a function not only of their tendency to produce good results but also of their usefulness in overcoming specific problems and limitations that we face. Many of these problems concern the cognitive and temporal limitations we human beings suffer from. Thinking through complex decision problems takes time and effort and can be undermined by factors like emotion and peer pressure. We are vulnerable to temptation, fickleness, sour grapes, and despair. Therefore, we should deliberate with norms that aid us with these challenges, and there is no guarantee that these can be identified *a priori* – empirical research has an important role to play here.

More controversially, we suggest that the context in which the agent normally operates, or reasonably expects to operate, also plays an important role (Morton 2016).[2] Agents in different contexts face different kinds of problems, and so should use different norms of reasoning to solve those problems. We have effectively already appealed to the relevance of context in discussing the problem of normative underdetermination. We have need of norms that help with the problem of underdetermination, but agents in a context where there is always a clear answer about what is best to do will not. Another example comes from reflecting on the very different limitations agents face due to socioeconomic circumstances. Agents with scarce resources tend to endure more burdens on their "cognitive bandwidth" and face higher pressure to solve short-term problems in the most resource-efficient way possible (Mullainathan and Shafir 2013). Therefore, it may well be adaptive for them to habitually reason in a way that prioritizes short-term goals over long-term ones, even if they deem the long-term ones more important. In contrast, agents with a more comfortable margin of error will likely do better to reason with more familiar norms favoring long-term prudence. Again, this is partly an empirical question.

We have focused in our examples on norms that favor defeasible diachronic stability in our intentions. What of the other norms we have been considering – the Instrumental Principle, Consistency, and Enkrasia? First, regularly violating these principles would render us extremely ineffective in bringing about our ends.[3] The skeptic replies that this is all to the good if you are an agent who tends to have bad ends, but most of us are not Iago or Caligula, and we will do far better if we are disposed not to step on our own feet. Second, having attitudes that are mutually coherent is highly useful for being interpretable to others, as well as oneself. Even if there are cases in which an agent will be more successful by having mutually inconsistent intentions, he will find it very difficult to coordinate with anyone else, or with himself over time. Kolodny (2008) denies that the value of interpretability helps to justify coherence as such, since the disposition to conform to one's reasons would serve this purpose just as well. But as we hope to have convinced the reader by now, this is not a disposition creatures like us can realistically just *manifest*. The fact that ideal reasons-responders need not care about coherence to be interpretable means little to us.

[2] More needs to be said about how to individuate the particular context. However, as we see it, the context should be individuated only as finely as the agent's cognitive capacities for detecting that change in context make it feasible to do so. A norm that is only suitable for a context that is more fine-grained than the agent can reasonably detect is not a very useful norm for that agent.

[3] Some might argue that inconsistency in one's attitudes is irrational even in worlds in which such inconsistency would make us more effective, for example, in a world in which a demon rewards such inconsistency. On our view, this assessment of irrationality would indeed be liable to a charge of rule-worship.

Many thanks to Kurt Sylvan, Matthew Silverstein, and Michael Bratman for comments on an earlier draft. This entry was written while receiving support from the John Templeton Foundation. The views expressed do not necessarily reflect the views of the Templeton Foundation.

Finally, what if we give up the metaethical assumption guiding the discussion in this section – that reasons for intention and action exist independently of the perspective of practical reason? We will close by floating the possibility of a Constitutivist account that embraces many of the features of the ecological account. One might be a "contingent constitutivist" who holds that the norms any particular agent should use are constitutive of some aspects of her identity but that these aspects need not be so universal as "agency as such." Rather, they might include much more specific features like her physical and cognitive limitations, her biological nature, and the social and cultural roles she inhabits. The thought would be that these specific ways of being an agent in the material and social world could generate norms that are contingent and therefore not materially inescapable, but that nonetheless have authority for her in virtue of the structural role that they play – they define for that agent how to think, and so changing them would be a radical move akin to a paradigm shift in science (Walden 2012). As on the ecological view, the right way to investigate these kinds, roles, functions, and the norms that apply to them would be at least partly empirical rather than purely *a priori*.

In the 1950s, Herbert Simon called for economists to replace

> the global rationality of economic man with a kind of rational behavior that is compatible with access to information and the computational capacities actually possessed by organisms, including man, in the kinds of environments in which such organisms exist.
>
> (1955, 99)

Simon saw this task as inherently interdisciplinary. And in the sixty years since, economists and psychologists have made great strides in understanding how human beings actually reason (Kahneman 2011; Gigerenzer and Goldstein 1996) and how the environment affects that capacity (Mullainathan and Shafir 2013). One benefit of our approach is that it offers a way of seeing the task of understanding practical deliberation as continuous with work done in other disciplines. The task is to offer a theory of deliberation for creatures like ourselves who, despite our cognitive limitations, adapt to a diversity of environments.

References

Arpaly, Nomy. 2003. *Unprincipled Virtue: An Inquiry into Moral Agency*. Oxford: Oxford University Press.

———. 2000. "On Acting Rationally Against One's Own Best Judgment." *Ethics* 110 (3): 488–513.

Arpaly, Nomy, and Timothy Schroeder. 2013. *In Praise of Desire*. Oxford: Oxford University Press.

———. 2012. "Deliberation and Acting for Reasons." *The Philosophical Review* 121 (2): 209–239.

Bratman, Michael. 1996. "Identification, Decision, and Treating as a Reason." *Philosophical Topics* 24 (2): 1–18.

———. 1987. *Intention, Plans, and Practical Reason*. Stanford: CSLI Publications.

Broome, John. 2013. *Rationality Through Reasoning*. West Sussex: Wiley-Blackwell.

———. 2007a. "Wide or Narrow Scope?" *Mind* 116 (462): 359–370.

———. 2007b. "Is Rationality Normative?" *Disputatio* 2 (23): 161–178.

———. 1999. "Normative Requirements." *Ratio* 12 (4): 398–419.

Carroll, Lewis. 1895. "What the Tortoise Said to Achilles." *Mind* 4 (14): 278–280.

Chang, Ruth. 2009. "Voluntarist Reasons and the Sources of Normativity." In *Reasons for Action*, edited by David Sobel and Steven Wall, 243–271. Cambridge: Cambridge University Press.

———. 2002. "The Possibility of Parity." *Ethics* 112 (4): 659–688.

Gigerenzer, Gerd and Goldstein, Daniel G. 1996. "Reasoning the Fast and Frugal Way: Models of Bounded Rationality." *Psychological Review* 103 (4): 650–669.

Harman, Gilbert. 1976. "Practical Reasoning." *The Review of Metaphysics* 29 (3): 431–463.

Holton, Richard. 2009. *Willing, Wanting, Waiting*. Oxford: Oxford University Press.

Kahneman, Daniel. 2011. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Kolodny, Niko. 2011. "Aims as Reasons." In *Reasons and Recognition: Essays on the Philosophy of T.M. Scanlon*, edited by Samuel Freeman, Rahul Kumar, and R. Jay Wallace, 43–78. Oxford: Oxford University Press.

———. 2008. "Why Be Disposed to Be Coherent?" *Ethics* 118 (3): 437–463.

———. 2007. "State or Process Requirements?" *Mind* 116 (462): 371–385.

———. 2005. "Why Be Rational?" *Mind* 114 (455): 509–563.

Korsgaard, Christine. 2008. *The Constitution of Agency*. Oxford: Oxford University Press.
Morton, Jennifer. 2016. "Reasoning Under Scarcity." *Australasian Journal of Philosophy* 95 (3): 543–559.
Morton, Jennifer, and Paul, Sarah. 2019. "Grit." *Ethics* 129 (2): 175–203.
Mullainathan, Sendhil, and Shafir, Eldar. 2013. *Scarcity: Why Having Too Little Means So Much*. New York: Henry Holt and Company.
Paul, Sarah. 2014. "Diachronic Incontinence Is a Problem in Moral Philosophy." *Inquiry* 57 (3): 337–355.
———. 2011. "Deviant Formal Causation." *Journal of Ethics & Social Philosophy* 5(3).
Railton, Peter. 1984. "Alienation, Consequentialism, and the Demands of Morality." *Philosophy & Public Affairs* 13 (2): 134–171.
Raz, Joseph. 2005. "The Myth of Instrumental Rationality." *Journal of Ethics and Social Philosophy* 1 (1).
———. 1999. "Incommensurability and Agency." In *Engaging Reason*, edited by Joseph Raz. Oxford: Oxford University Press.
Scanlon, Thomas. 2004. "Reasons: A Puzzling Duality?" In *Reason and Value. Themes from the Moral Philosophy of Joseph Raz*. Oxford: Oxford University Press, 231–246.
Setiya, Kieran. 2007. "Cognitivism About Instrumental Reason." *Ethics* 117 (4): 649–673.
Simon, Herbert. 1955. "A Behavioral Model of Rational Choice." *The Quarterly Journal of Economics* 69 (1): 99–118.
Smart, J. J. C. 1956. "Extreme and Restricted Utilitarianism." *The Philosophical Quarterly* 6 (25): 344–354.
Tiffany, Evan. 2012. "Why Be an Agent?" *Australasian Journal of Philosophy* 90 (2): 223–233.
Velleman, J. David. 2009. *How We Get Along*. Cambridge: Cambridge University Press.
———. 2007. "What Good Is a Will?" In *Action in Context*, edited by Anton Leist and Holger Baumann. Berlin: de Gruyter/Mouton.
Walden, Kenneth. 2012. "Laws of Nature, Laws of Freedom, and the Social Construction of Normativity." In *Oxford Studies in Metaethics Volume 7*. Oxford: Oxford University Press.
Wallace, R. Jay. 2001. "Normativity, Commitment, and Instrumental Reason." *Philosophers' Imprint* 1 (3).
Way, Jonathan. 2011. "The Symmetry of Rational Requirements." *Philosophical Studies* 155 (2): 227–239.
Wood, Allen. 1990. *Hegel's Ethical Thought*. Cambridge: Cambridge University Press.

Appendix: A Guide to Further Reading

Preliminary remarks

As the Introduction emphasized, the literature on practical reason is a quicksand with unclear boundaries. Inevitably, then, there are (i) unexamined angles on topics that we have covered in the volume, (ii) topics that we have not covered, and (iii) adjacent literatures that might be classified as philosophy of practical reason, but which we have mostly bracketed. This Appendix provides a guide to further reading on (i) and (ii).

But first, a few more words about (iii). For the same reason that it is worth distinguishing between moral philosophy and the philosophy of practical reason, it is worth distinguishing between the *theory of prudence* and the philosophy of practical reason. There are prudential reasons for action, to be sure. But the study of these reasons is inseparable from the study of well-being and the good life and hence from ethics broadly construed. The reason-implying status of ethics is contested. A significant tradition – Humeanism – denies that either prudence or morality gives normative reasons as such, independently of our present desires. Although Humeanism may be false, it is more helpful for the introductory reader to appreciate this distinction and note that, for some philosophers of practical reason, prudence is as open to question as morality. Some literature before the 1990s is misleading about this point. For example, under the influence of Sidgwick, Parfit's first book treated some questions within the theory of prudence as questions about practical reason.[1] For explorations of these questions in the spirit of Parfit (1984), we suggest McMahan (2002), Brink (1997a-1997b), Crisp (2006), and Schechtman's (2014) discussions of prudential concern, personal identity, and well-being; Sidgwick (1874) is key background. For more recent work at the intersection with the theory of prudence (and decision theory), see Hedden (2015) and Sullivan (2018).

A similar distinction seems worth drawing between (much of) decision theory and the philosophy of practical reason. Decision theory seeks to provide a formal model of rational preference and choice. Parts of decision theory consist in the mathematical and logical study of properties of this model rather than in the philosophical defense of the adequacy of the model. These parts are related to the philosophy of practical reason in roughly the way that probability theory is related to epistemology. Probability theory is not a branch of epistemology, though the philosophical study of certain interpretations of probability (e.g., the Bayesian interpretation) is a branch of formal epistemology.

[1] See Parts II and III of Parfit (1984) and then compare the different take on rationality in Parfit (2011).

But even here it might be worth drawing a distinction. Unless one is a specific kind of consequentialist about normative reasons, one will not want to take it for granted that expected utility theory is an *explanatory* theory of practical normativity. Even if one thought rational choice had to be *representable* by expected utility theory – which is questionable – it would not follow that rational choice is at any level *explained* by utilitarian principles. If 'consequentializers' are to be believed (see Dreier 1993), any normative theory will have an extensionally equivalent consequentialist counterpart. But it does not follow that consequentialism is a true explanatory theory. Yet the normative parts of the philosophy of practical reason seek explanatory theories of reasons and rationality. Hence, just as the study of consequentialism could only be a small fraction of ethics, even the philosophical parts of decision theory could only be a small fraction of the philosophy of practical reason.

Nevertheless, a student of the philosophy of practical reason should get to know decision theory, for the same reason that a student of epistemology should get to know Bayesianism. For overviews with significant contributions, see Buchak (2013) and Steele and Stefansson (2015); for some classic foundational works, see Jeffrey (1965) and Joyce (1999). Some important formally sophisticated works in the philosophy of practical reason include Broome (1991), Sen (1977, 1987), Sepielli (2009, 2014), Temkin (2012), and Wedgwood (2017).

1 Further historical reading

Let's turn to our guide to further reading. To begin with, there is much more reading worth doing on practical reason in the history of philosophy, both on the figures and traditions we have covered and on traditions that aren't covered in the book.

1.1 More readings on traditions and figures covered

Let's first consider more resources on what we have covered.

Classical Chinese Philosophy. As Wong's title indicates, his focus in the paper is on early Chinese philosophy (6^{th} century BCE to 221 BCE), in particular on the Confucian and Daoist traditions. There are other periods and traditions to consider. For a more complete overview, Wong's entry on Chinese ethics in the *Stanford Encyclopedia of Philosophy* is worth consulting (2018); while it covers ethical matters that go beyond the philosophy of practical reason, it has an emphasis throughout on the practical character of Chinese ethics and serves as a useful companion to his entry.

Aristotle. The literature on Aristotle's philosophy of practical reason is divided. Some philosophers (e.g., Kolnai 1962) assume Aristotle is an *instrumentalist* because of a remark in *Nichomachean Ethics* (1112b11–12) often translated as saying that practical deliberation is of means, not ends. Others push back and argue that Aristotle acknowledges what Millgram (2001) calls 'specificationist' reasoning, which is a form of deliberation that moves from wider to narrower ends; see especially Wiggins (1975). For further discussion of how to interpret Aristotle here, see Cooper (1986), McDowell (1979), Nussbaum (1986: Ch.10), Price (2011a-2011b), and Sorabji (1980). Other important topics treated by Aristotle include the 'guise of the good' account of motivation and the nature of weakness of will; on these topics, see Dahl (1984), Kenny (1979), and Moss (2010, 2012). For a full overview of both sides of the literature, see Taylor (2016).

Hume. Sayre-McCord's interpretation of Hume is heterodox: as he mentions, it is more standard to see Hume as being either a skeptic about practical reason or an instrumentalist. But Sayre-McCord has some fellow travelers, including Dorsey (2008), Schafer (2016), and Setiya (2004). For the skeptical interpretation, see Millgram (1995), Hampton (1995), and Korsgaard (1986). The instrumentalist interpretation is pervasive among philosophers of practical reason who aren't primarily focused on exegesis. This fact is responsible for the popular use of 'Humean' to mean 'instrumentalist'.

Kant. Since a whole handbook could be devoted to Kant's philosophy of practical reason, the introductory reader is strongly advised to read beyond our volume. Schapiro takes a novel angle on Kant in this volume, adopting a high-level comparative focus that doesn't get too lost in the exegetical trenches. This is a great way to get introduced to Kant's approach to practical reason. But it is also worth getting acquainted with other angles on Kant, like those involving close exegesis and those involving an equal balance between exegesis and philosophical defense. For the latter, the work of Christine Korsgaard is compulsory for all students of the philosophy of practical reason; see especially Korsgaard (1996a, 1996b). The work of Onora O'Neill is similarly important; see especially O'Neill (1975, 1989, 2004). O'Neill's work has received less attention in the wider philosophy of practical reason than Korsgaard's, but it merits equal recognition and is at least as influential in Kantian circles. For close exegetical work, see Reath and Timmerman (2010); for a broader overview, see Wood (2013). Recent exegetical literature has emphasized that Kant's philosophy of practical reason should be understood in light of a unity claim he makes, according to which theoretical and practical reason are uses of 'one and the same reason' (see *Critique of Practical Reason* 5:121); for an overview, see Williams (2017), and for some contributions, see Neiman (1994) and Kleingeld (1998). In reflecting on this topic, one must also take into account Kant's further thought that practical reason has primacy; see Mudd (2016) for a discussion of this topic which also contributes to the previous literature.

Anscombe. Anscombe is celebrated equally for some separable contributions to the philosophy of practical reason. Singh's contribution focuses on the issues surrounding her well-known claim that intentional action is action which gives application to a certain sense of the question 'Why?' This part of Anscombe secures her an important place in the history of the philosophy of practical reason and not just the history of action theory, since it suggests that the theory of intentional action is part of the first branch of the philosophy of practical reason. But Anscombe also had an influential picture of practical reasoning as 'calculative' in form, which isn't discussed at length in Singh's contribution. The reader is advised to look at the literature that discusses or is inspired by this theory. Here we recommend Schwenkler (2019), Vogler (2002), and Wiseman (2016: Ch.5). As Müller (2011) notes, ascription of this theory to Anscombe may rest on negligence of her discussion of 'backward-looking reasons' (though she admittedly doesn't develop this idea much in *Intention*).

1.2 Readings on traditions and figures not covered

Practical reason appears elsewhere in the history of philosophy, including both 'Western' and 'non-Western' philosophy. In the case of Western philosophy, insights into practical reason can be found in Plato, some Hellenistic philosophy, medieval philosophy, the rationalists, some 18th-century British philosophers besides Hume, 19th-

and early 20th-century European and British philosophy, and Continental philosophy. In the case of non-Western philosophy, underappreciated insights can be found in classical Indian philosophy, medieval Islamic philosophy (itself continuous with medieval European philosophy), and Africana philosophy. A full guide to each of these periods and traditions is beyond the scope of this Appendix. But to do some justice to them, we will direct the reader to some resources which seemed like good starting points for a student of practical reason in history.

Our selection from the history of Western philosophy represents the figures that have been especially influential on contemporary philosophy of practical reason; as we noted previously, views labeled 'Humean', 'Kantian', 'Aristotelian', or 'Anscombean' dominate the literature. But historians of philosophy have found views within the philosophy of practical reason in other figures and periods.

First, a quick run from ancient Greek philosophy to medieval philosophy: In the case of Plato, some important recent studies include Barney (2010), Evans (2010), Kamtekar (2017), and Price (2011c). Frede (1994) and Brennan (2003) cover some relevant issues in the Stoics; the intersection between practical reason and Epicureanism is rarely explored, but Frugé (2020) argues that Epicureanism leads to skepticism about practical reason; finally, Perin (2010) discusses the intersection of Pyrrhonian skepticism with the first and third branches. Celano (2018) provides a dedicated overview of medieval work on practical reason; Tenenbaum (2007a) discusses some of the medieval background to the guise of the good while defending the doctrine.

Now, a jog from rationalism to late 19th- and early 20th-century British philosophy. For discussions of the rationalists on the relationship between reason and emotion, see Gaukroger (ed.) (1998). For Descartes on action, motivation, and reason, see Greenberg (2007), Kenny (1972), Naaman-Zauderer (2010), and Williston (1999). Studies of Spinoza on action, motivation, and reason can be found in Della Rocca (2003), Hübner (2018), Kisner (2011), Marshall (2013), Naaman-Zauderer (ed.) (2020), and Rutherford (2008). Schapiro (2001) discusses the intersection of action theory and the philosophy of practical reason in a comparison of the neglected 17th-century utilitarian Richard Cumberland and neglected 17th-century deontologist William Wollaston. Darwall (1995) covers the internalism vs. externalism debate in British philosophy between 1640 and 1740. For an overview of Locke's philosophy of action, see Rickless (2020). Finally, Reid's *Essays on the Active Powers of Man (1788/2010*) presents his philosophy of action, and this provides one of the few cases in pre-20th century history of a dedicated treatise on action theory. The 19th-century in British philosophy then sees the pivotal figure of Sidgwick, whose approach anticipates (through its influence on Parfit) the contemporary second branch of the philosophy of practical reason discussed in the Introduction; in Hurka (2014), one can see an overview of relevant late 19th- and early 20th-century work in the second branch. Mill was less influential than Sidgwick on the philosophy of practical reason, but he has received some discussion under this heading; see Millgram (2000) and Skorupski (1989: Chs.8–9).

To round off Western philosophy, let's consider Continental philosophy, which contains a trove of writings on practical reason that analytic philosophers have largely (but not entirely) ignored. Hegel's conception of practical reason is clearly explained and examined in Kain (1998), Pinkard (1994), Pippin (2008), Skorupski (2010), and Wood (1990). Lessons for the philosophy of practical reason can be extracted from Marx in several ways. On the one hand, as Wood (1999) and Hurka have (1993) noted, Marx combined Aristotelian constitutivism with a novel conception of human nature to derive reasons for rejecting capitalism. Commentators outside analytic philosophy, such as Harvey (2014, 2017) and Marcuse (1964), have also seen an attempt in Marx's discussion of the 'contradictions' of capitalism to expose its *irrationality*; the style of argument here is reminiscent of the Kantian constitutivist's objections to immorality. One can also find in Marx some objections to an instrumentalist model of rationality. This point is developed best in the Frankfurt school, by Max Horkheimer (1947, 2012). Related ideas appear in Arendt (1958: Part IV) and in Weber's discussion of the eclipse of 'value rationality' (*Wertrationalität*) by instrumental rationality (*Zweckrationalität*) in Weber (1921/1968); see Brubaker (1984) for a discussion of Weber of particular interest for the philosophy of practical reason. Two other areas of Continental philosophy that have recently been considered for their insights into the philosophy of practical reason are Nietzsche and existentialism. Katsafanas (2013) argues that a novel form of constitutivism can be derived from some of Nietzsche's remarks, Ridley (2008) argues that the kind of 'expressivist' model of agency discussed in Schapiro (2001) can be found in a distinctive form in Nietzsche, and the essays in Janaway and Robertson (2012) consider Nietzsche from related angles. Webber (2018) does for French existentialism what Katsafanas did for Nietzsche. Heidegger is approached in a related spirit in Crowell (2013), McManus (2015), and Rousse (2016).

We conclude with non-Western philosophy, which has been just as neglected as Continental philosophy. We will start with classical Indian philosophy, which dates back at least to the *Upaniṣads* and the Pali canon of Buddhism (c. 800–500 BCE and c. 100 BCE, respectively), then consider classical Islamic philosophy and Africana philosophy.

While the contributions of classical India to epistemology are increasingly recognized and have a long history of study, similar claims are less clear for the philosophy of practical reason. Discussions of action theory, moral psychology, and the place of rationality in the Brahmanical systems (*darśanas*) that were developed through *sūtras* and commentaries can be found in Matilal (2007), Mohanty (2007), and Raghunathan (2017); a discussion of relevant issues in Indian literature can be found in Mathur (1974) and Matilal (2002). Discussions of action and ethics in Indian Buddhism can be found in Heim (2013), in the essays by Siderits, Meyers, and Repetti in Davis (2017), and in many essays in Repetti (2016). Framarin (2009) and Chuang (2015) consider the idea of desireless action in classical Indian literature and philosophy. Broader overviews of the central place of rationality in classical Indian philosophy can be found in Ganeri (2001) and Mohanty (1992). Finally, it is worth adding that a striking feature of both

Brahmanical and Buddhist systems is their emphasis on the role of *knowledge* in proper action, and it could be argued that the idea of a *knowledge norm* on action was not a novelty of 21st-century work like Hawthorne and Stanley (2008) but rather a staple of classical Indian philosophy; see, for example, Vātsyāyana's commentary on *Nyāya-Sūtra* 1.1.1 and the opening of Dharmottara's *Nyāyabindu*.[1]

Much less literature can be found relevant to the study of practical reason in Islamic philosophy and Africana philosophy. In the case of Islamic philosophy, relevant discussions can be found in Al-Ghazali and Ibn-Rushd (Averroes), and there is undoubtedly more to be discovered in the reception of Aristotle in classical Arabic philosophy. It remains insufficiently known that Al-Ghazali (c.1056–1111) presented a case with the same structure and purpose as Burdian's ass more than two centuries earlier; see Marmura (2002: 23) for the passage. He also sought to draw some significant lessons about the limits of reason from the example. Averroes defended reason from the example with some points about the individuation of objects of choice; see McGinnis and Reisman (2007: 304). Averroes also penned a significant discussion of weakness of will and played an important role in the reception of Aristotle's philosophy of action and mind in medieval Latin philosophy; see Saarinen (1994: Ch.3, Sect.1) and Phillipson (2013, 2017) on these matters.

Even less has been done in the case of African and Africana philosophy. Two areas that merit further attention concern the existence in Akan philosophy of the person of a constitutivist account of morality in terms of the constitution of personhood, as well as the existence of a reasons-responsive view of responsibility; on these matters, see Wingo (2006) and Wiredu (1992), respectively. In the history of African-American philosophy, a figure who stands out as especially in need of attention in the philosophy of practical reason is Alain Locke, who wrote important works on value conflict, incommensurability, and value pluralism; see Carter (2012: §§2–3).

[1] Vātsyāyana (translated by Dasti and Phillips 2017): 'When an object is comprehended through a knowledge source, it becomes possible to engage in successful goal-directed activity. Thus, a knowledge source is useful. Without a knowledge source, there would be no effective cognition of an object. Without such cognition, there would be no successful action. When someone grasps something by means of a knowledge source, the person may desire to obtain or to avoid it. *Goal-directed activity* is the effort of someone who acts because of such a desire or aversion'; Dharmottara (translated by Stcherbatsky (1962)): 'All successful human action is preceded by knowledge.' For a fuller discussion of the idea in Vātsyāyana, see Dasti (2017).

2 Further reading at the intersection with action theory and moral psychology

2.1 More readings on aspects of topics covered

An important region of literature near Arpaly's and Buss's contributions that the reader should get to know concerns *reasons-responsiveness* accounts of free will and moral responsibility. Classic defenses of this view can be found in Fischer and Ravizza (1998), Nelkin (2011), Wolf (1994), and Smith (2003b). For an overview of the literature on this topic which separately discusses Wolf's and Nelkin's views and Fischer and Ravizza's views, see McKenna and Coates (2019: §§4.3–4.4).

For further discussion of the conclusion of practical reasoning, see Anscombe (1957: §33), Audi (1989), Tenenbaum (2007b), and Paul (2013). For an overview, see Streumer (2010). A central connected question also at the heart of Dancy's contribution is the relationship between theoretical and practical reasoning. Cognitivists about practical reason seek to reduce practical to theoretical reasoning. Brunero (2014) and Lord (2018a) serve as useful critical introductions to the topic.[1] The main defenders of this view are Harman (1976), Setiya (2007), Velleman (1989), and Wallace (2001). Bratman is a persistent critic; see, for example, Bratman (2009a). New light has been shed on this topic in Fix (2018), who explores a wider disagreement between intellectualists (including cognitivists) and opponents of intellectualism about practical reason, and who opposes intellectualism and defends the distinctive practicality of practical reason. This framing of the issues suggests an interesting link with the literature on know-how (see Pavese (2016a-2016b) for an overview), a link which can also be found in Setiya (2016).

A short list of further work we would recommend on the guise of the good includes classic papers by Stocker (1979), Raz (2002: Ch.2) and Velleman (1992), Orsi's (2015) survey article, and everything in Tenenbaum's (2010) collection. For another overview

[1] Note that cognitivism about practical reason is a different view from the more familiar kind of cognitivism defended by Scanlon. In discussing Scanlon in the Introduction, we used 'cognitivism' in its commonest sense to mean the view that normative judgments express beliefs. But cognitivism about practical reason is the different view that practical reason is grounded in or is a special application of theoretical reason.

of work on motivational internalism vs. externalism, see Björklund et al. (2012); for a significant collection of recent work, see Björnsson et al. (2015). Some must-read discussions of motivational internalism's role in metaethics include Smith (1994) and Svavarsdóttir (1999); Smith's (1991) survey article on realism compresses some of the relevant points in Smith (1994). Finally, Bagnoli (2011) is a nice collection on emotions and moral psychology; some relevant monographs besides Greenspan (1993) are Helm (2001), Roberts (2003), and Tappolet (2016).

2.2 Readings on topics not covered

Because the first branch of the philosophy of practical reason almost always overlaps with two other large areas (action theory and philosophy of mind), it is a challenge to include everything relevant without veering too much into the other areas. Much work produced in Davidsonian action theory could be placed in the first branch. But it is more plausible to classify such work as *primarily* action theory, which is one reason we did not survey it in the Introduction. For two series of papers on Davidson's action theory, see Lepore and McLaughlin (1985) and Part 1 of Lepore and Ludwig (2013). For a basic overview, see Chapter 4 of Glüer (2011). Wiland (2012: Ch. 2) provides an especially relevant critical discussion of Davidsonian philosophy of action. For a comprehensive reader in action theory which is especially relevant for students of philosophy of practical reason, see Dancy and Sandis (2015).

Besides work on the relationship between reason and freedom, a student of philosophy of practical reason should get to know the literature on the relationship between *acting for a reason* and *acting intentionally*. Anscombe (1957) can be credited with this idea, but the idea pervades the Davidsonian tradition as well. Mele (1992) is a neglected Davidsonian paper which defends the identification of acting for a reason and acting intentionally. The identification has been questioned by experimental philosophy; see Knobe and Kelly (2009) for an experimental attack. Hursthouse (1991) is a classic armchair work opposing the identification.

The student of the philosophy of practical reason must also get to know the Humean theory of motivation, which has its 20$^{\text{th}}$-century origins in Davidson. For an overview, see Nottelmann (2011). For some classic work in favor of it, see Smith (1987), who also takes stock of the most important challenges to the view at the time he was writing. Schueler (1995) and Shafer-Landau (2003) offer some classic discussions after Smith of the limits of the view. For a full-throated recent defense, see Sinhababu (2017). For a defense of the role of desire in motivation, see Arpaly and Schroeder (2014).

Davidsonians and fans of the Humean theory of motivation typically accept a psychological ontology of motivating reasons, according to which motivating reasons *consist in* beliefs and desires. But this view took a dramatic hit in popularity among philosophers of reason after Dancy (2000), who argued that it made it impossible to act for normative reasons. This argument stimulated new research on the ontology of

motivating reasons, which is another area it would have been nice to cover. For an overview of the literature which seeks to accommodate the intuitions behind both psychological and non-psychological ontologies, see Alvarez (2010). Mitova (2017) gives an impressive defense of psychologism that turns Dancy's argument on its head in the course of defending a new ontology of epistemic reasons.

Followers of Dancy will still accept certain psychological constraints on motivation but just argue that the reasons by which we are motivated are not psychological states. One important question that has received a lot of recent attention concerns the *epistemic* relation we must bear to a reason in order to act for that reason. Hyman (2015) argues that it is the knowledge relation, while Dancy rejects this view and argues that we can act for the reason that p even if it is not the case that p; see Dancy (2011) for a critical discussion of earlier papers by Hyman. Some other important work under this heading includes Hornsby (2008) and Locke (2015).

A final topic which came up throughout the book to which we could have devoted a whole volume is the nature of practical reasoning and the varieties it can take. Millgram (2001) remains an outstanding reader on this topic, though one should now supplement it with Broome (2013b). Two paper-length introductions are Streumer (2010) and Kauppinen (2018). Two older monographs that seem worthy of a revival are Audi (1989) and Richardson (1994).

3 Further reading on practical normativity

This volume's section on the metaphysics of reasons seems to us sufficiently complete that we will only make a few recommendations under this heading. But there is much more that could have been covered in the second section of Part 4, so we will spend more time on it.

3.1 More readings on aspects of topics covered

Section 1 Topics. Paakkunainen and Setiya (2011) collects much of the most important work on subjectivist and Kantian accounts of the metaphysics of reasons; Chapters 5 and 6 of Mele and Rawling (2004) and Chapters 6 and 11 of Star (2018) also cover this ground. Wiland (2012) provides a great book-length introduction to metaphysical questions about both normative and motivating reasons, which covers all the major accounts (including hybrid theories). Finlay and Schroeder (2017) gives an article-length introduction, and Section 2 of Alvarez (2016) offers a section-length introduction. The *loci classici* of Humeanism, Kantianism, Aristotelianism, general constitutivism, hybrid voluntarism, naturalist realism, and non-naturalism, respectively, are: Schroeder (2007a) and Sobel (2017), Korsgaard (1996a, 1996b, 2009) and O'Neill (1989), Foot (2001) and Thomson (2008), Velleman (2000) and Smith (2013), Chang (2009, 2013), Railton (1984), Parfit (2011, 2017), Scanlon (1998, 2013), and Skorupski (2010).

Section 2 Topics. Lord and Maguire (2016) collects recent work on the weight of reasons. Chang (1997) collects some classic papers on comparability, commensurability, and practical reason. Baumann and Betzler (2004) collects work on practical dilemmas. Anderson (1993) and Stocker (1990) are classic book-length discussions of pluralism and practical reason. Crisp (2018) discusses the conflict of prudential and moral reasons. Dancy (1993), Darwall (1983, 2006), Korsgaard (1996b), and Markovits (2014) are classic book-length discussions of moral normativity.

3.2 Readings on topics not covered

Further Part 1 Topics. The contributions in the volume mainly concern the metaphysical relation between the normative and the non-normative. But there are impor-

tant metaphysical questions within the normative domain: we can ask which normative properties or relations are *most fundamental* and how the reason-relation might itself be analyzed in other normative terms.

Reasons-firsters argue that all normativity is grounded in reasons. For defenses of this view, see Schroeder (2007a, forthcoming), Skorupski (2010), and Lord (2018b). For more general discussions and criticisms, see Part III of Star (2018). There are also more specific questions about how to explain certain normative properties in terms of reasons; see Scanlon (1998: Ch.2), Rabinowicz and Ronnow-Rasmussen (2004), Schroeder (2010, 2016: Section 3), Jacobson (2011), and Rowland (2019) for some important discussions of how to ground value in reasons. That literature gave rise to a more general literature on how to distinguish between reasons which are of the 'right kind' for analyzing normative properties and reasons of the 'wrong kind'; for attempts to solve this problem, see Hieronymi (2005), Lord and Sylvan (2019), and Schroeder (2010). Reasons-based accounts generated pushback by other philosophers seeking to analyze reasons in terms of values, virtues, or oughts; for such views, see Maguire (2016), Thomson (2008), Broome (2004), and Kearns and Star (2009).

Further Part 2 Topics. Part 2 was an expandable ragbag for stray topics. There are many other topics it would have been nice to cover with dedicated entries if we had more space, such as:

> A The distinction between agent-neutral and agent-relative reasons;
>
> B The distinction between reasons one possesses and reasons that may be unpossessed;
>
> C The nature of acting for good reasons (or correctly responding to reasons);
>
> D Epistemic norms on action and practical norms on belief;
>
> E The debate between satisficing and maximizing views about practical reason;
>
> F Holism and particularism about normative practical reasons;
>
> G Teleological vs. non-teleological theories of normative practical reasons.

The list could go on and on. Since this Appendix cannot go on too much longer, let's focus on a brief guide to further reading on each of these issues.

> (A) Ridge (2005) and Bykvist (2018) provide overviews of the literature on agent-neutral and agent-relative reasons. Some classic general discussions of this topic are Nagel (1970, 1986: Ch. 9), Parfit (1984: Parts 1 and 2), Scheffler (1982), and Sen (1983). For a discussion of the agent-relative and agent-neutral debate about relationship-based reasons, see Jeske (2008) and Jollimore (2001). See Smith (2003a), Portmore (2005), and Schroeder

(2007b) for an examination of whether consequentialism can or should be stated in an agent-relative form.

(B) The concept of a normative reason is used in two ways: we can think about the reasons we *possess* (where this possession involves at least epistemic access), and we can also think about reasons in abstraction from our epistemic access to them. The literature on the relationship between these uses of the concept of a reason has expanded greatly in recent years. For some theories of what it is to possess a reason, see Vogelstein (2012), Whiting (2014), Sylvan (2015), Lord (2018b), and Wodak (2019). For a broader overview of work on this distinction, see Sepielli (2018). Note that this distinction is often called the distinction between 'subjective' and 'objective' normative reasons. But the subjective/objective distinction at issue here is usually taken to be orthogonal to the distinction that divides the theorists Parfit calls 'Subjectivists' and 'Objectivists'.

(C) Another literature that has recently bloomed examines the nature of *correctly responding* to reasons. It is possible to be moved by a consideration that *happens* to be a good reason without manifesting sensitivity to the normative relation between that consideration and one's action or intention. The point is familiar in epistemology: one might reason from some premises that in fact entail a conclusion, but still reason fallaciously in doing so, by using a bad rule. As Lord (2018b) and Mantel (2018) argue, the same goes in ethics, and we must hence reject the view that acting for good reasons is just a matter of having motivating reasons which *coincide* with good reasons (cf. Dancy 2000 and Markovits 2010). What such sensitivity involves is a contested matter. Arpaly (2003) and Harman (2011) argue that it isn't a matter of having a rational belief that one's reason for acting is a good reason. But Johnson King (forthcoming) makes some points which help to resist them.

(D) A third literature which intersects with the previous two focuses on the idea that there are *epistemic norms* on action. This literature emerged from epistemology as part of the defense of *pragmatic encroachment*, which holds that knowledge and epistemic normativity are partly grounded in pragmatic factors. Fantl and McGrath (2002, 2009) and Hawthorne and Stanley (2008) defended this view on the basis of the principle that one should act on an assumption only if one knows this assumption to be the case. This principle, in turn, was defended on the basis of the intuitive thought that it is objectionable for a person to act on an assumption without knowing it to be true. This literature partly inspired the study of the intersection of practical and epistemic reason. While epistemologists have been more interested in whether epistemology is 'encroached upon' by pragmatic reasons, their key arguments have, in effect, relied on the assumption

that the pragmatic is encroached upon by the epistemic. But not everyone buys this claim; see Simion (2018) for a critical discussion.
While the literature on pragmatic encroachment is best regarded as part of epistemology, it is part of a broader literature which students of practical reason should study: namely the literature on whether practical reason might govern *belief* as well as action and intention. Several philosophers of practical reason have been doubtful about this idea; see, for example, Parfit's (2011) appendix on 'state-given' reasons and Skorupski's (2010) discussion of the distinction between epistemic and practical reasons. But some epistemologists have continued to defend pragmatic reasons for belief; if they were right, then the philosophy of practical reason would cover some norms on belief, not just norms on action and intention; see Leary (2017), McCormick (2015), and Rinard (2015).

(E) Some readers will be familiar with the distinction between maximizing consequentialism, which obliges us to maximize goodness, and satisficing consequentialism, which obliges us to bring about good enough consequences. But the distinction between these theories is a special case of a more general debate about practical reason. An independent philosophical literature developed on that broader debate, which has also been explored outside philosophy (e.g., in economic anthropology (see Sahlins 1974) and behavioral economics (see Schwartz et al. 2010). Byron (2004) is an outstanding collection on the topic. Slote (1989) and Stocker (1990) are classic defenses of satisficing. Pettit (1984) and Bradley (2006) offer important critiques.

(F) Although Dancy's contribution to this volume contains some discussion of holism and particularism about reasons, it is not his main focus. For book-length defenses of both views by Dancy, see his (1993) and (2000). Other defenses of particularism can be found in Lance and Little (2006) and McNaughton and Rawling (2000). Arguments against particularism can be found in Berker (2007) and McKeever and Ridge (2006). For an article overviewing the issues, see Väyrynen (2018). For separate discussions of holism, see Schroeder (2011) and Bader (2016).

(G) A final literature we mentioned in passing earlier is on the status of *teleological* theories of reasons. According to teleological theories, all normative reasons for action are explained by means of some connection to the *promotion of value*. Portmore (2011a-2011b) and Maguire (2016) defend this view. Anderson (1993), Hurley (2017, 2019), Raz (2011, 2016) and Scanlon (1998: Ch.2) argue against it, with Raz defending the strongly non-teleological view that *no* reasons are reasons to promote value.

4 Further reading on practical rationality

We end with some further readings for the third branch of the philosophy of practical reason.

4.1 More readings on aspects of topics covered

Lord's contribution to the volume covers two issues that can be distinguished: the *scope* of coherence requirements and their *normative status*. For further reading on scope, see Broome (2007), Brunero (2010), Lord (2014), Shpall (2013), Schroeder (2004), Way (2010a, 2011), and Worsnip (2015). For an alternative overview of the literature on the normativity of rationality, see Way (2010b); for central monographs, see Kiesewetter (2017), Lord (2018b), and Wedgwood (2017). Further discussions of instrumental rationality in particular can be found in Brunero (2020), Kolodny and Brunero (2018), and Way (2012, 2013).

New literature has emerged which is dedicated to the question of whether (as Bratman believes) there are diachronic requirements of rationality. Ferrero (2012, 2014) argues that there are, while Hedden (2015) and Paul (2014) are important opponents. Although Moss (2015) is interested in epistemology, her arguments for 'time-slice epistemology' have significant bearing on this debate.

New literature has also emerged that examines the relationship between rational requirements and reasoning, which is a topic in the background of Morton and Paul's paper. Hussain (MS) is an influential unpublished work which defended the idea that rational requirements govern reasoning before Broome (2013b) began to develop the idea at length (though it was implicit in Broome 1999). McHugh and Way (2015, 2018) critique Broome's view and develop a different account of the standards of correctness for reasoning.

4.2 Readings on topics not covered

Considerable bodies of writing have emerged on more specific rational requirements and forms of irrationality. For example, there has been much recent discussion of Broome's Enkrasia principle; see the special issue of *Organon F* that contains Broome

(2013a) and Reisner (2013), and also see Coates (2013), Hinchman (2013), and Way (forthcoming). Some insights have come from epistemologists interested in higher-order evidence; see especially Lasonen-Aarnio (2020).

There is an older literature about weakness of will. Stroud and Tappolet (2003) is a classic collection; Stroud and Svirsky (2019) give a new overview. Much of the earlier literature equated weakness of will with acting or intending against one's better judgment. But Holton (2009) argued that we should distinguish weakness of will and akrasia. For a book-length study by a frequent contributor to the topic that was written after Holton which defends the equation, see Mele (2012).

References

Alvarez, M. 2010. Kinds of Reasons: An Essay in the Philosophy of Action. *Oxford: Oxford University Press.*

Alvarez, M. 2016. 'Reasons for Action: Justification, Motivation, and Explanation' in *Stanford Encyclopedia of Philosophy.* https://plato.stanford.edu/entries/reasons-just-vs-expl/.

Anderson, E. 1993. *Value in Ethics and Economics.* Cambridge, MA: Harvard University Press.

Anscombe, E. 1957. *Intention.* Oxford: Basil Blackwell.

Arendt, H. 1958. *The Human Condition.* Chicago: University of Chicago Press.

Arpaly, N. 2003. *Unprincipled Virtue.* Oxford: Oxford University Press.

Arpaly, N. and Schroeder, T. 2014. *In Praise of Desire.* Oxford: Oxford University Press.

Audi, R. 1989. *Practical Reasoning.* Abingdon: Routledge.

Bader, R. 2016. 'Conditions, Modifiers, and Holism' in Lord, E. and Maguire, B. (eds.) *Weighing Reasons.* New York: Oxford University Press.

Bagnoli, C. (ed.) 2011. *Morality and the Emotions.* Oxford: Oxford University Press.

Barney, R. 2010. 'Plato on the Desire for the Good' in Tenenbaum, S. (ed.) *Desire, Practical Reason, and the Good.* Oxford: Oxford University Press.

Baumann, P. and Betzler, M. (eds.) 2004. *Practical Conflicts.* Cambridge: Cambridge University Press.

Berker, S. 2007. 'Particular Reasons' *Ethics* 118: 109–139.

Björklund, F. (et al.) 2012. 'Recent Work on Motivational Internalism' *Analysis* 72: 124–137.

Björnsson, G. (et al.) 2016. *Motivational Internalism.* Oxford: Oxford University Press.

Bradley, B. 2006. 'Against Satisficing Consequentialism' *Utilitas* 17: 282–298.

Bratman, M. 2009a. 'Intention, Belief, and Practical Rationality' in Sobel, D. and Wall, S. (eds.) *Reasons for Action.* Cambridge: Cambridge University Press.

Brennan, T. 2003. 'Stoic Moral Psychology' in Inwood, B. (ed.) *The Cambridge Companion to the Stoics.* Cambridge: Cambridge University Press.

Brink, D. O. 1997a. 'Rational Egoism and the Separateness of Persons' in Dancy, J. (ed.) *Reading Parfit.* Oxford: Blackwell.

Brink, D. O. 1997b. 'Self-Love and Altruism' *Social Philosophy and Policy* 14: 122–157.

Broome, J. 1991. *Weighing Goods.* Oxford: Blackwell.

Broome, J. 1999. 'Normative Requirements.' *Ratio* 12: 398–419.
Broome, J. 2004. 'Reasons' in Wallace, R. J., Smith, M., Scheffler, S. and Pettit, P. (eds.) *Reason and Value: Themes from the Philosophy of Joseph Raz*. Oxford: Oxford University Press.
Broome, J. 2007. 'Wide or Narrow Scope?' *Mind* 116: 359–370.
Broome, J. 2013a. 'Enkrasia' *Organon F* 20: 425–436.
Broome, J. 2013b. *Rationality Through Reasoning*. Oxford: Blackwell.
Brubaker, R. 1984. *The Limits of Rationality*. Abingdon: Routledge.
Brunero, J. 2010. 'The Scope of Rational Requirements' *Philosophical Quarterly* 60: 28–49.
Brunero, J. 2014. 'Cognitivism About Practical Rationality' *Oxford Studies in Metaethics* 9: 18–44.
Brunero, J. 2020. *Instrumental Rationality*. Oxford: Oxford University Press.
Buchak, L. 2013. *Risk and Rationality*. Oxford: Oxford University Press.
Bykvist, K. 2018. 'Agent-Relative and Agent-Neutral Reasons' in Star, D. (ed.) *The Oxford Handbook of Reasons and Normativity*. Oxford: Oxford University Press.
Byron, M. (ed.) 2004. Satisficing and Maximising: Moral Theorists on Practical Reason. *Cambridge: Cambridge University Press.*
Carter, J. A. 2012. 'Alain LeRoy Locke' in *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/alain-locke/.
Celano, A. 2018. 'Medieval Theories of Practical Reason' in *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/practical-reason-med/.
Chang, R. (ed.) 1997. *Incommensurability, Incomparability, and Practical Reason*. Cambridge, MA: Harvard University Press.
Chang, R. 2009. 'Voluntarist Reasons and the Sources of Normativity' in Sobel, D. and Wall, S. (eds.) *Reasons for Action*. Cambridge: Cambridge University Press.
Chang, R. 2013. 'Grounding Practical Normativity: Going Hybrid.' *Philosophical Studies* 164: 164–187.
Chuang, C. 2015. 'Understanding Desireless Action as Benevolent Action' *Asian Philosophy* 25: 132–147.
Coates, A. 2013. 'The Enkratic Requirement' *European Journal of Philosophy* 21: 320–333.
Cooper, J. 1986. *Reason and Human Good in Aristotle*. Indianapolis: Hackett.
Crisp, R. 2006. *Reasons and the Good*. Oxford: Oxford University Press.
Crisp, R. 2018. 'Prudential and Moral Reasons' in Star, D. (ed.) *The Oxford Handbook of Reasons and Normativity*. Oxford: Oxford University Press.
Crowell, S. 2013. *Normativity and Phenomenology in Husserl and Heidegger*. Cambridge: Cambridge University Press.
Dahl, N. O. 1984. *Practical Reason, Aristotle, and Weakness of Will*. Minneapolis: University of Minnesota Press.
Dancy, J. 1993. *Moral Reasons*. Oxford: Blackwell.
Dancy, J. 2000. *Practical Reality*. Oxford: Oxford University Press.

Dancy, J. 2011. 'Acting in Ignorance' *Frontiers of Philosophy in China* 6: 345–357.
Dancy, J. and Sandis, C. 2015. *Philosophy of Action: An Anthology*. Oxford: Blackwell.
Darwall, S. 1983. *Impartial Reason*. Ithaca: Cornell University Press.
Darwall, S. 1995. *The British Moralists and the Internal 'Ought'*. Cambridge: Cambridge University Press.
Darwall, S. 2006. *The Second-Person Standpoint*. Cambridge, MA: Harvard University Press.
Dasti, M. 2017. 'Vātsyāyana on Cognition as a Guide to Action' in Ganeri, J. (ed.) *The Oxford Handbook of Indian Philosophy*. Oxford: Oxford University Press.
Dasti, M. and Phillips, S. (transl.) 2017. *The Nyāya-Sūtra*. Indianapolis: Hackett.
Davis, J. H. (ed.) 2017. *A Mirror Is for Reflection: Understanding Buddhist Ethics*. Oxford: Oxford University Press.
Della Rocca, M. 2003. 'The Power of an Idea: Spinoza's Critique of Pure Will' *Nous* 37: 200–231.
Dorsey, D. 2008. 'Hume's Internalism Reconsidered' *The Journal of Ethics and Social Philosophy* 2(3): 1–23.
Dreier, J. 1993. 'Structures of Normative Theories' *The Monist* 76: 22–40.
Evans, M. 2010. 'A Partisan's Guide to Socratic Intellectualism' in Tenenbaum, S. (ed.) *Desire, Practical Reason, and the Good*. Oxford: Oxford University Press.
Fantl, J. and McGrath, M. 2002. 'Evidence, Pragmatics and Justification.' *Philosophical Review* 111: 67–94.
Fantl, J. and McGrath, M. 2009. *Knowledge in an Uncertain World*. Oxford: Oxford University Press.
Ferrero, L. 2012. 'Diachronic Constraints of Practical Rationality' *Philosophical Issues* 22: 144–164.
Ferrero, L. 2014. 'Diachronic Structural Rationality' *Inquiry* 57: 311–336.
Finlay, S. and Schroeder, M. 2017. 'Reasons for Action: Internal and External.' *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/reasons-internal-external/.
Fischer, J. M. and Ravizza, M. 1998. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.
Fix, J. D. 2018. 'Intellectual Isolation' *Mind* 506: 491–520.
Foot, P. 2001. *Natural Goodness*. Oxford: Oxford University Press.
Framarin, C. G. 2009. *Desire and Motivation in Indian Philosophy*. Abingdon: Routledge.
Frede, M. 1994. 'The Stoic Conception of Reason' in Bourdouris, K. (ed.) *Hellenistic Philosophy* (vol. 2). Athens: International Center for Greek Philosophy and Culture.
Frugé, C. 2020. 'Epicureanism and Skepticism About Practical Reason' *Canadian Journal of Philosophy* 50: 195–208.
Ganeri, J. 2001. *Classical Indian Philosophy*. Abingdon: Routledge.
Gaukroger, S. (ed.) 1998. The Soft Underbelly of Reason: The Passions in the 18th Century*. London: Routledge.*

Glüer, K. 2011. *Donald Davidson: A Very Short Introduction*. Oxford: Oxford University Press.
Greenberg, S. 2007. 'Descartes on the Passions: Function, Representation, and Motivation' *Nous* 41: 714–734.
Greenspan, P. 1993. *Emotions and Reasons*. London: Routledge.
Hampton, J. 1995. 'Does Hume Have an Instrumental Conception of Practical Reason?' *Hume Studies* 21: 57–74.
Harman, E. 2011. 'Does Moral Ignorance Exculpate?' *Ratio* 24: 443–468.
Harman, G. 1976. 'Practical Reasoning' *Review of Metaphysics* 29: 431–463.
Harvey, D. 2014. Seventeen Contradictions and the End of Capitalism. *London: Profile Books.*
Harvey, D. 2017. Marx, Capital, and the Madness of Economic Reason. *London: Profile Books.*
Hawthorne, J. and Stanley, J. 2008. 'Knowledge and Action' *Journal of Philosophy* 105: 571–590.
Hedden, B. 2015. *Reasons Without Persons*. Oxford: Oxford University Press.
Heim, M. 2013. The Forerunner of All Things: Buddhaghosa on Mind, Intention, and Agency. *Oxford: Oxford University Press.*
Helm, B. 2001. *Emotional Reason*. Cambridge: Cambridge University Press.
Hieronymi, P. 2005. 'The Wrong Kind of Reason' *Journal of Philosophy* 102: 437–457.
Hinchman, E. 2013. 'Rational Requirements and "Rational" Akrasia' *Philosophical Studies* 166: 520–552.
Holton, R. 2009. *Willing, Wanting, Waiting*. Oxford: Oxford University Press.
Horkheimer, M. 1947. *The Eclipse of Reason*. Oxford: Oxford University Press.
Horkheimer, M. 2012. *The Critique of Instrumental Reason*. London: Verso Books.
Hornsby, J. 2008. 'A Disjunctive Conception of Acting for Reasons' in Haddock, A. and MacPherson, F. (eds.) *Disjunctivism: Perception, Action, Knowledge*. Oxford: Oxford University Press.
Hübner, K. 2018. 'Spinoza's Unorthodox Metaphysics of Will' in Della Rocca, M. (ed.) *The Oxford Handbook of Spinoza*. Oxford: Oxford University Press.
Hurka, T. 1993. *Perfectionism*. Oxford: Oxford University Press.
Hurka, T. 2014. *British Ethical Theorists from Sidgwick to Ewing*. Oxford: Oxford University Press.
Hurley, P. 2017. 'Consequentialism and the Standard Story of Action' *Journal of Ethics* 22: 22–44.
Hurley, P. 2019. 'Exiting the Consequentialist Circle: Two Senses of "Bringing About"' *Analytic Philosophy* 60: 130–163.
Hursthouse, R. 1991. 'Arational Actions' *Journal of Philosophy* 88: 57–68.
Hussain, N. MS. 'The Requirements of Rationality' Unpublished manuscript, Stanford University.
Hyman, J. 2015. *Action, Knowledge, and Will*. Oxford: Oxford University Press.

Jacobson, D. 2011. 'Fitting Attitudes Accounts of Value' in *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/fitting-attitude-theories/.
Janaway, C. and Robertson, S. (eds.) 2012. *Nietzsche and Normativity*. Oxford: Oxford University Press.
Jeffrey, R. C. 1965. *The Logic of Decision*. New York: McGraw-Hill.
Jeske, D. 2008. Rationality and Moral Theory: How Intimacy Generates Reasons. *Abingdon: Routledge.*
Johnson King, Z. Forthcoming. 'Accidentally Doing the Right Thing.' *Philosophy and Phenomenological Research*.
Jollimore, T. 2001. *Friendship and Agent-Relative Morality*. Abingdon: Routledge.
Joyce, J. 1999. *The Foundations of Causal Decision Theory*. Cambridge: Cambridge University Press.
Kain, P. 1998. 'Hegel's Critique of Kantian Practical Reason' *Canadian Journal of Philosophy* 28: 367–412.
Kamtekar, R. 2017. *Plato's Moral Psychology*. Oxford: Oxford University Press.
Katsafanas, P. 2013. Agency and the Foundations of Ethics: Nietzschean Constitutivism. *Oxford: Oxford University Press.*
Kauppinen, A. 2019. 'Practical Reasoning' in Star, D. (ed.). *The Oxford Handbook of Reasons and Normativity*. Oxford: Oxford University Press.
Kearns, S. and Star, D. 2009. 'Reasons as Evidence' *Oxford Studies in Metaethics* 4: 415–442.
Kenny, A. 1972. 'Descartes on the Will' in Butler, R. J. (ed.) *Cartesian Studies*. Oxford: Oxford University Press.
Kenny, A. 1979. *Aristotle's Theory of the Will*. London: Duckworth.
Kiesewetter, B. 2017. *The Normativity of Rationality*. Oxford: Oxford University Press.
Kisner, M. 2011. Spinoza on Human Freedom: Reason, Autonomy, and the Good Life. *Cambridge: Cambridge University Press.*
Kleingeld, P. 1998. 'Kant on the Unity of Theoretical and Practical Reason' *The Review of Metaphysics* 52: 311–339.
Knobe, J. and Kelly, S. D. 2009. 'Can One Act for a Reason Without Acting Intentionally?' In Sandis, C. (ed.) *New Essays on the Explanation of Action*. Basingstoke: Palgrave Macmillan.
Kolnai, A. 1962. 'Deliberation Is of Ends' *Proceedings of the Aristotelian Society* 62: 195–218.
Kolodny, N. and Brunero, J. 2018. 'Instrumental Rationality' in *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/rationality-instrumental/.
Korsgaard, C. 1986. 'Skepticism about Practical Reason' *Journal of Philosophy* 83: 5–25.
Korsgaard, C. 1996a. *Creating the Kingdom of Ends*. Cambridge: Cambridge University Press.
Korsgaard, C. 1996b. *The Sources of Normativity*. Cambridge: Cambridge University Press.

Korsgaard, C. 2009. *Self-Constitution*. Oxford: Oxford University Press.
Lance, M. and Little, M. O. 2006. 'Defending Moral Particularism' in Dreier, J. (ed.) *Contemporary Debates in Moral Theory*. Oxford: Blackwell.
Lasonen-Aarnio, M. 2020. 'Enkrasia or Evidentialism? Learning to Love Mismatch' *Philosophical Studies* 177: 597–632.
Leary, S. 2017. 'In Defense of Practical Reasons for Belief' *Australasian Journal of Philosophy* 95: 529–542.
Lepore, E. and Ludwig, K. (eds.) 2013. *A Companion to Donald Davidson*. Oxford: Blackwell.
Lepore, E. and McLaughlin, B. (eds.) 1985. Actions and Events: Perspectives on the Philosophy of Donald Davidson. *Oxford: Blackwell.*
Locke, D. 2015. 'Knowledge, Explanation, and Motivating Reasons' *American Philosophical Quarterly* 52: 215–232.
Lord, E. 2014. 'The Real Symmetry Problem(s) for Wide-Scope Accounts of Rationality' *Philosophical Studies* 170: 443–464.
Lord, E. 2018a. 'The Explanatory Problem for Cognitivism About Practical Reason' in McHugh, C., Way, J. and Whiting, D. (eds.) *Normativity: Epistemic and Practical*. Oxford: Oxford University Press.
Lord, E. 2018b. *The Importance of Being Rational*. Oxford: Oxford University Press.
Lord, E. and Maguire, B. (eds.) 2016. *Weighing Reasons*. Oxford: Oxford University Press.
Lord, E. and Sylvan, K. 2019. 'Reasons: Wrong, Right, Normative, Fundamental' *Journal of Ethics and Social Philosophy* 15: 43–74.
Maguire, B. 2016. 'The Value-Based Theory of Reasons' *Ergo* 3: 233–262.
Mantel, S. 2018. *Determined by Reasons*. London: Routledge.
Marcuse, H. 1964. *One-Dimensional Man*. London: Routledge.
Markovits, J. 2010. 'Acting for the Right Reasons' *Philosophical Review* 119: 201–242.
Markovits, J. 2014. *Moral Reason*. Oxford: Oxford University Press.
Marmura, M. E. (transl.) 2002. *Al-Ghazali's The Incoherence of the Philosophers*. Chicago: University of Chicago Press.
Marshall, E. 2013. The Spiritual Automaton: Spinoza's Science of the Mind. *Oxford: Oxford University Press.*
Mathur, D. C. 1974. 'The Concept of Action in Bhagavad-Gita' *Philosophy and Phenomenological Research* 35: 34–45.
Matilal, B. K. 2002. *Ethics and Epics*. Oxford: Oxford University Press.
Matilal, B. K. 2007. '*Dharma* and Rationality' in Bilimoria, P., Prabhu, J. and Sharma, R. (eds.) *Indian Ethics* (vol. 1). Aldershot: Ashgate.
McCormick, M. S. 2015. *Believing Against the Evidence*. Abingdon: Routledge.
McDowell, J. 1979. 'Virtue and Reason' *The Monist* 62: 331–350.
McGinnis, J. and Reisman, D. C. 2007. *Classical Arabic Philosophy: An Anthology of Sources*. Indianapolis: Hackett.
McHugh, C. and Way, J. 2015. 'Broome on Reasoning' *Teorema* 34: 131–140.

McHugh, C. and Way, J. 2018. 'What Is Good Reasoning?' *Philosophy and Phenomenological Research* 94: 153–174.
McKeever, S. and Ridge, M. 2006. *Principled Ethics*. Oxford: Clarendon Press.
McKenna, M. and Coates, J. D. 2019. 'Compatibilism' in *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/compatibilism/.
McMahan, J. 2002. *The Ethics of Killing*. Oxford: Oxford University Press.
McManus, D. (ed.) 2015. *Heidegger, Authenticity, and the Self*. Abingdon: Routledge.
McNaughton, D. and Rawling, P. 2000. 'Unprincipled Ethics' in Hooker, B. and Little, M. O. (eds.) *Moral Particularism*. Oxford: Clarendon Press.
Mele, A. 1992. 'Acting for Reasons and Acting Intentionally' *Pacific Philosophical Quarterly* 73: 355–374.
Mele, A. 2012. *Backsliding: Understanding Weakness of Will*. Oxford: Oxford University Press.
Mele, A. and Rawling, P. (eds.) 2004. *The Oxford Handbook of Rationality*. Oxford: Oxford University Press.
Millgram, E. 1995. 'Was Hume a Humean?' *Hume Studies* 21: 75–93.
Millgram, E. 2000. 'Mill's Proof of the Principle of Utility' *Ethics* 110: 282–310.
Millgram, E. (ed.) 2001. *The Varieties of Practical Reasoning*. Cambridge, MA: The MIT Press.
Mitova, V. 2017. *Believable Evidence*. Cambridge: Cambridge University Press.
Mohanty, J. N. 1992. *Reason and Tradition in Indian Thought*. Oxford: Oxford University Press.
Mohanty, J. N. 2007. '*Dharma*, Imperatives, and Tradition: Toward an Indian Theory of Moral Action' in Bilimoria, P., Prabhu, J. and Sharma, R. (eds.) *Indian Ethics* (vol. 1). Aldershot: Ashgate.
Moss, J. 2010. 'Aristotle's Non-Trivial, Non-Insane View That Everyone Always Desires Things Under the Guise of the Good' in Tenenbaum, S. (ed.) *Desire, Practical Reason, and the Good*. Oxford: Oxford University Press.
Moss, J. 2012. *Aristotle on the Apparent Good*. Oxford: Oxford University Press.
Moss, S. 2015. 'Time-Slice Epistemology and Action Under Indeterminacy' *Oxford Studies in Epistemology* 5: 172–194.
Mudd, S. 2016. 'Rethinking the Priority of Practical Reason in Kant' *European Journal of Philosophy* 24: 78–102.
Müller, A. 2011. 'Backward-Looking Rationality and the Unity of Practical Reason' in Ford, A., Hornsby, J. and Stoutland, F. (eds.) *Essays on Anscombe's Intention*. Cambridge, MA: Harvard University Press.
Naaman-Zauderer, N. 2010. *Descartes's Deontological Turn*. Cambridge: Cambridge University Press.
Naaman-Zauderer, N. (ed.) 2020. *Freedom, Action, and Motivation in Spinoza's Ethics*. London: Routledge.
Nagel, T. 1970. *The Possibility of Altruism*. Princeton: Princeton University Press.
Nagel, T. 1986. *The View from Nowhere*. Oxford: Oxford University Press.

Neiman, S. 1994. *The Unity of Reason*. Oxford: Oxford University Press.
Nelkin, D. 2011. *Making Sense of Freedom and Responsibility*. Oxford: Oxford University Press.
Nottelmann, N. 2011. 'Belief-Desire Explanation' *Philosophy Compass* 6: 912–921.
Nussbaum, M. 1986. *The Fragility of Goodness*. Cambridge: Cambridge University Press.
O'Neill, O. 1975. *Acting on Principle*. Cambridge: Cambridge University Press.
O'Neill, O. 1989. *Constructions of Reason*. Cambridge: Cambridge University Press.
O'Neill, O. 2004. 'Kant: Rationality as Practical Reason' in Mele, A. and Rawling, P. (eds.) *The Oxford Handbook of Rationality*. Oxford: Oxford University Press.
Orsi, F. 2015. 'The Guise of the Good' *Philosophy Compass* 10: 714–724.
Paakkunainen, H. and Setiya, K. (eds.) 2011. *Internal Reasons: Contemporary Readings*. Cambridge, MA: The MIT Press.
Parfit, D. 1984. *Reasons and Persons*. Oxford: Clarendon Press.
Parfit, D. 2011. *On What Matters* (vol. 1). Oxford: Oxford University Press.
Parfit, D. 2017. *On What Matters* (vol. 3). Oxford: Oxford University Press.
Paul, S. K. 2013. 'The Conclusion of Practical Reasoning: Between Shadow and Act' *Canadian Journal of Philosophy* 43: 287–302.
Paul, S. K. 2014. 'Diachronic Incontinence Is a Problem in Moral Philosophy' *Inquiry* 57: 337–355.
Pavese, C. 2016a. 'Skill in Epistemology I: Skill and Knowledge' *Philosophy Compass* 11: 642–649.
Pavese, C. 2016b. 'Skill in Epistemology II: Skill and Know-How' *Philosophy Compass* 11: 650–660.
Perin, C. 2010. *The Demands of Reason*. Oxford: Oxford University Press.
Pettit, P. 1984. 'Satisficing Consequentialism' *Proceedings of the Aristotelian Society* 58: 165–176.
Phillipson, T. 2013. 'The Will in Averroes and Aquinas' *Proceedings of the American Catholic Association* 87: 231–247.
Phillipson, T. 2017. *Aquinas, Averroes, and the Human Will*. PhD Thesis, Marquette University, Milwaukee.
Pinkard, T. 1994. *Hegel's Phenomenology: The Sociality of Reason*. Cambridge: Cambridge University Press.
Pippin, R. 2008. Hegel's Practical Philosophy: Rational Agency as Ethical Life. *Cambridge: Cambridge University Press.*
Portmore, D. 2005. 'Combining Teleological Ethics with Evaluator Relativism: A Promising Result' *Philosophical Quarterly* 86: 95–113.
Portmore, D. 2011a. *Commonsense Consequentialism*. Oxford: Oxford University Press.
Portmore, D. 2011b. 'The Teleological Conception of Practical Reasons' *Mind* 120: 117–153.

Price, A. W. 2011a. 'Aristotle on Practical Reasoning' in *Virtue and Reason in Plato and Aristotle*. Oxford: Oxford University Press.
Price, A. W. 2011b. 'Aristotle on the Ends of Deliberation' in Pakaluk, M. and Pearson, G. (eds.) *Moral Psychology and Human Action in Aristotle*. Oxford: Oxford University Press.
Price, A. W. 2011c. *Virtue and Reason in Plato and Aristotle*. Oxford: Oxford Uniersity Press.
Rabinowicz, W. and Ronnow-Rasmussen, T. 2004. 'The Strike of the Demon: On Fitting Pro-Attitudes and Value' *Ethics* 114: 391–423.
Raghunathan, R. 2017. 'Two Theories of Motivation and Their Assessment by Jayanta' in Ganeri, J. (ed.) *The Oxford Handbook of Indian Philosophy*. Oxford: Oxford University Press.
Railton, P. 1984. 'Moral Realism' *Philosophical Review* 95: 163–207.
Raz, J. 2002. *Engaging Reason*. Oxford: Oxford University Press.
Raz, J. 2011. From Normativity to Responsibility. Oxford: Oxford University Press.
Raz, J. 2016. 'Value and the Weight of Reasons' in Lord, E. and Maguire, B. (eds.) *Weighing Reasons*. New York: Oxford University Press.
Reath, A. and Timmerman, J. (eds.) 2010. *Kant's Critique of Practical Reason: A Critical Guide*. Cambridge: Cambridge University Press.
Reid, T. 1788/2010. *Essays on the Active Powers of Man*. Edinburgh: University of Edinburgh Press.
Reisner, A. 2013. 'Is the Enkratic Principle a Requirement of Rationality?' *Organon F* 20: 436–462.
Repetti, R. (ed.) 2016. Buddhist Perspectives on Free Will: Agentless Agency? *Abingdon: Routledge.*
Richardson, H. 1994. *Practical Reasoning about Final Ends*. Cambridge: Cambridge University Press.
Rickless, S. 2020. 'Locke on Freedom' in *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/locke-freedom/.
Ridge, M. 2005. 'Reasons for Action: Agent-Neutral vs. Agent-Relative.' *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/reasons-agent/.
Ridley, A. 2008. *The Deed Is Everything*. Oxford: Oxford University Press.
Rinard, S. 2015. 'Against the New Evidentialists' *Philosophical Issues* 25: 208–223.
Roberts, R. C. 2003. *Emotions: An Essay in Aid of Moral Psychology*. Cambridge: Cambridge University Press.
Rousse, B. S. 2016. 'Heidegger, Sociality, and Human Agency' *European Journal of Philosophy* 24: 417–451.
Rowland, R. 2019. *The Normative and the Evaluative*. Oxford: Oxford University Press.
Rutherford, D. 2008. 'Spinoza on the Dictates of Reason' *Inquiry* 5: 485–511.
Saarinen, R. 1994. Weakness of the Will in Medieval Thought. *Leiden: Brill.*
Sahlins, M. 1974. *Stone Age Economics*. Abingdon: Routledge.

Scanlon, T. M. 1998. *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.
Scanlon, T. M. 2013. Being Realistic about Reasons. Oxford: Oxford University Press.
Schafer, K. 2016. 'Hume on Practical Reason' in Russell, P. (ed.) *The Oxford Handbook of Hume*. Oxford: Oxford University Press.
Schapiro, T. 2001. 'Three Conceptions of Action in Moral Theory' *Nous* 35: 93–117.
Schechtman, M. 2014. Staying Alive: Personal Identity, Practical Concerns, and the Unity of a Life. *Oxford: Oxford University Press.*
Scheffler, S. 1982. *The Rejection of Consequentialism*. Oxford: Oxford University Press.
Schroeder, M. 2004. 'The Scope of Instrumental Reason' *Philosophical Perspectives* 18: 337–364.
Schroeder, M. 2007a. *Slaves of the Passions*. Oxford: Oxford University Press.
Schroeder, M. 2007b. 'Teleology, Agent-Relative Value and "Good"' *Ethics* 116: 265–295.
Schroeder, M. 2010. 'Value and the Right Kind of Reason' in Shafer-Landau, R. (ed.) *Oxford Studies in Metaethics 5*. Oxford: Oxford University Press.
Schroeder, M. 2011. 'Holism, Weight and Undercutting' *Nous* 45: 328–344.
Schroeder, M. 2016. 'Value Theory' *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/value-theory/.
Schroeder, M. Forthcoming. *Reasons First*. Oxford: Oxford University Press.
Schueler, G. F. 1995. Desire: Its Role in Practical Reason and the Explanation of Action. *Cambridge, MA: The MIT Press.*
Schwartz, B., Ben-Haim, Y. and Dacso, C. 2010. 'What Makes a Good Decision? Robust Satisficing as a Normative Standard of Rational Decision Making' *Journal for the Theory of Social Behaviour* 41: 209–227.
Schwenkler, J. 2019. *A Guide to Anscombe's Intention*. Oxford: Oxford University Press.
Sen, A. 1977. 'Rational Fools: A Critique of the Behavioral Foundations of Economic Theory' *Philosophy and Public Affairs* 6: 317–344.
Sen, A. 1983. 'Evaluator Relativity and Consequential Evaluation' *Philosophy and Public Affairs* 12: 113–132.
Sen, A. 1987. *On Ethics and Economics*. Oxford: Blackwell.
Sepielli, A. 2009. 'What to Do When You Don't Know What to Do' in Shafer-Landau, R. (ed.) *Oxford Studies in Metaethics 4*. Oxford: Oxford University Press.
Sepielli, A. 2014. 'What to Do When You Don't Know What to Do What You Don't Know What to Do ...' *Nous* 48: 521–544.
Sepielli, A. 2018. 'Subjective and Objective Reasons' in Star, D. (ed.) *The Oxford Handbook of Reasons and Normativity*. Oxford: Oxford University Press.
Setiya, K. 2004. 'Hume on Practical Reason' *Philosophical Perspectives* 18: 365–389.
Setiya, K. 2007. 'Cognitivism about Instrumental Reason' *Ethics* 117: 649–673.
Setiya, K. 2016. *Practical Knowledge*. Oxford: Oxford University Press.
Shafer-Landau, R. 2003. *Moral Realism: A Defense*. Oxford: Oxford University Press.

Shpall, S. 2013. 'Wide and Narrow Scope' *Philosophical Studies* 163: 717–736.
Sidgwick, H. 1874. *The Methods of Ethics*. London: MacMillan and Co.
Simion, M. 2018. 'No Epistemic Norm for Action' *American Philosophical Quarterly* 55: 231–238.
Sinhababu, N. 2017. *Humean Nature*. Oxford: Oxford University Press.
Skorupski, J. 1989. *John Stuart Mill*. Abingdon: Routledge.
Skorupski, J. 2010. *The Domain of Reasons*. Oxford: Oxford University Press.
Slote, M. 1989. *Beyond Optimizing*. Cambridge, MA: Harvard University Press.
Smith, M. 1987. 'The Humean Theory of Motivation' *Mind* 96: 36–61.
Smith, M. 1991. 'Realism' in Singer, P. (ed.) *A Companion to Ethics*. Oxford: Blackwell.
Smith, M. 1994. *The Moral Problem*. Oxford: Blackwell.
Smith, M. 2003a. 'Neutral and Relative Value After Moore' *Ethics* 113: 576–598.
Smith, M. 2003b. 'Rational Capacities' in Stroud, S. and Tappolet, C. (eds.) *Weakness of Will and Practical Irrationality*. Oxford: Clarendon Press.
Smith, M. 2013. 'A Constitutivist Theory of Reason: Its Promise and Parts' *Law, Ethics, and Philosophy* 1: 1–30.
Star, D. (ed.) 2018. *The Oxford Handbook of Reasons and Normativity*. Oxford: Oxford University Press.
Sobel, D. 2017. *From Valuing to Value*. Oxford: Oxford University Press.
Sorabji, R. 1980. 'Aristotle on the Role of Intellect in Virtue' in Rorty, A. O. (ed.) *Essays in Aristotle's Ethics*. Berkeley: University of California Press.
Stcherbatsky, F. 1962. *Buddhist Logic* (vol. 2). New York: Dover.
Steele, K. and Stefansson, H. O. 2015. 'Decision Theory' in *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/decision-theory/.
Stocker, M. 1979. 'Desiring the Bad: An Essay in Moral Psychology' *Journal of Philosophy* 76: 739–753.
Stocker, M. 1990. *Plural and Conflicting Values*. Oxford: Clarendon Press.
Streumer, B. 2010. 'Practical Reasoning' in O'Connor, T. and Sandis, C. (eds.) *Blackwell Companion to the Philosophy of Action*. Oxford: Blackwell.
Stroud, S. and Svirsky, L. 2019. 'Weakness of Will' in *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/weakness-will/.
Stroud, S. and Tappolet, C. (eds.) 2003. *Weakness of Will and Practical Irrationality*. Oxford: Clarendon Press.
Sullivan, M. 2018. Time Biases: A Theory of Rational Planning and Personal Persistence. *Oxford: Oxford University Press.*
Svavarsdóttir, S. 1999. 'Moral Cognitivism and Motivation' *Philosophical Review* 108: 161–219.
Sylvan, K. 2015. 'What Apparent Reasons Appear to Be' *Philosophical Studies* 172: 587–606.
Taylor, C. C. W. 2016. 'Aristotle on Practical Reason' in *Oxford Handbooks Online*. DOI: 10.1093/oxfordhb/9780199935314.013.52.
Tappolet, C. 2016. *Emotions, Values, and Agency*. Oxford: Oxford University Press.

Temkin, L. 2012. *Rethinking the Good*. Oxford: Oxford University Press.
Tenenbaum, S. 2007a. *Appearances of the Good*. Cambridge: Cambridge University Press.
Tenenbaum, S. 2007b. 'The Conclusion of Practical Reason' in Tenenbaum, S. (ed.) *New Trends in Philosophy: Moral Psychology*. Amsterdam: Rodopi.
Tenenbaum, S. (ed.) 2010. *Desire, Practical Reason, and the Good*. Oxford: Oxford University Press.
Thomson, J. J. 2008. *Normativity*. Chicago: Open Court.
Väyrynen, P. 2018. 'Reasons and Moral Principles' in Star, D. (ed.) *The Oxford Handbook of Reasons and Normativity*. Oxford: Oxford University Press.
Velleman, J. D. 1989. *Practical Reflection*. Princeton: Princeton University Press.
Velleman, J. D. 1992. 'The Guise of the Good' *Nous* 26: 3–26.
Velleman, J. D. 2000. *The Possibility of Practical Reason*. Oxford: Oxford University Press.
Vogler, C. 2002. *Reasonably Vicious*. Cambridge, MA: Harvard University Press.
Vogelstein, E. 2012. 'Subjective Reasons' *Ethical Theory and Moral Practice* 15: 239–257.
Wallace, R. J. 2001. 'Normativity, Commitment and Instrumental Reason' *Philosophers' Imprint* 1: 1–26.
Way, J. 2010a. 'Defending the Wide Scope Approach to Instrumental Reason' *Philosophical Studies* 147: 213–233.
Way, J. 2010b. 'The Normativity of Rationality' *Philosophy Compass* 5: 1057–1068.
Way, J. 2011. 'The Symmetry of Rational Requirements' *Philosophical Studies* 115: 227–239.
Way, J. 2012. 'Explaining the Instrumental Principle' *Australasian Journal of Philosophy* 90: 487–506.
Way, J. 2013. 'Instrumental Rationality' in Crane, T. (ed.) *Routledge Encyclopedia of Philosophy Online*. Abingdon: Routledge.
Way, J. Forthcoming. 'A Puzzle About Enkratic Reasoning' *Philosophical Studies*.
Webber, J. 2018. *Rethinking Existentialism*. Oxford: Oxford University Press.
Weber, M. 1921/1968. *Economy and Society*. New York: Bedminster Press.
Wedgwood, R. 2017. *The Value of Rationality*. Oxford: Oxford University Press.
Whiting, D. 2014. 'Keep Things in Perspective: Reasons, Rationality, and the a Priori' *Journal of Ethics and Social Philosophy* 8: 1–22.
Wiggins, D. 1975. 'Deliberation and Practical Reason' *Proceedings of the Aristotelian Society* 76: 29–51.
Wiland, E. 2012. *Reasons*. London: Continuum Press.
Williams, G. 2017. 'Kant's Account of Reason' in *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/kant-reason/.
Williston, B. 1999. 'Akrasia and the Passions in Descartes' *British Journal for the History of Philosophy* 7: 33–55.

Wingo, A. 2006. 'Akan Philosophy of the Person' in *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/akan-person/.
Wiredu, K. 1992. 'The Moral Foundation of an African Culture' in Flack, H. E. and Pellegrino, E. D. (eds.) *African-American Perspectives on Biomedical Ethics*. Washington, DC: Georgetown University Press.
Wiseman, R. 2016. The Routledge Guidebook to Anscombe's Intention*. London: Routledge.*
Wodak, D. 2019. 'An Objectivist's Guide to Subjective Reasons' *Res Philosophica* 96: 229–244.
Wolf, S. 1994. *Freedom Within Reason*. Oxford: Oxford University Press.
Wong, D. 2018. 'Chinese Ethics' in *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/ethics-chinese/.
Wood, A. 1990. *Hegel's Ethical Thought*. Cambridge: Cambridge University Press.
Wood, A. 1999. *Karl Marx*. London: Routledge.
Wood, A. 2013. 'Kant on Practical Reason' in Timmons, M. and Baiasu, S. (eds.) *Kant on Practical Justification*. Oxford: Oxford University Press.
Worsnip, A. 2015. 'Narrow-Scoping for Wide-Scopers' *Synthese* 192: 2617–2646.

Index

acting for reasons 10–11, 172; Anscombe's view on 179–182; and causes 174–177; and internal action 173–174; reasons as constituents of action 177–179
action 281–284; explaining actions done for reasons 357–359; and practical irrationality 197–200; reasons as constituents of 177–179
action, philosophy of 2–3
activist view 106
advice 445–446
affect 251–253; *see also* emotions
agency 262, 264–271, 272n9, 273n17, 273n19, 273n23; and Kant 168–169; and social science research 281–284; *see also* agency, theory of; human agency, theory of; morally responsible agency; passive agency
agency, theory of 165–166; dogmas of 187–194; *see also* human agency, theory of
agony 312–314
agony argument 300–306
akrasia 231–232
alienation 488–489
all things considered 130, 444, 446–447, 449–452, 454n11–12, 455n20
alternative constitutive aims 232–233
analogical reasoning 115–117
Anderson, Elizabeth xii, 9, 99, 496, 560
Andreou, Chrisoula xii, 16–17
Anscombe, G.E.M. 172, 179–182; intentional action and acting for reasons 173–174; reasons and causes 174–177; reasons as constituents of action 177–179
Aristotle 58–60, 126–127, 136–138; backwards reasoning 131–133; geometrical model for deliberation 127–131; on practical reason 216–217; and vice 133–136
Arpaly, Nomy xii, 11, 188, 545–546, 558–559, 561
attitudes 31–34
authority 463–464
backwards reasoning 131–133
bias 285–287
Bratman, Michael xii, 16–17, 100, 161–164, 558
Broome, John xii, 8–9, 15–17, 350, 471–475, 542–543
Buss, Sarah ii, 11, 558
Callard, Agnes xii, 9–10
canonical examples 487–488
case studies: passive agency 201–206; pragmatist methods 87–90
categorical rational requirements 542–544
causes 174–177
Chang, Ruth xii, 9, 12
choice *see* resolute choice; sophisticated choice; wise choice
choice over time 505–511

choice situations 103–106
Clarke 161–162
co-extensiveness argument 377–378
coherence 470–471; normative impotence of coherence requirements 472–473
commendatory reasons 398–400
concepts (term) 373–376
concrete experience 114–115
conditional internalism 239–241
conflict situations 420–421
Confucianism 123; contextual reasoning in 113–114
constitutivism 336–337, 345–346; and equivocations 343–344; methodology 345; and the Open Question Argument 342–343; and Shmagency 339–342; and Smith 337–339
constitutivist supplements 326–328
constructivism 318; constitutivist supplements 326–328; constructing respect for persons 322–326; constructing universalizability requirements 321–322; motivations for normative constructivism 318–320; reasoning and other people 328–333; varieties of normative constructivism 320–321
context 546–551; Anscombe's view on 179–182
contextual reasoning 113–114
contingency 546–551
Copp, David xii, 15, 240, 425
culture 72–74
Dancy, Jonathan xii, 11–12, 350, 353, 358, 405; and non-requiring reasons 393–394, 398–399
Daoism 123
Darwall, Stephen xii, 15; and second-personal reasons 460–463
deliberation 126–127, 136–138, 188–189; backwards reasoning 131–132; geometrical model for 127–131; and vice 133–136; *see also* practical deliberation
desire 117–118
directions of fit 60–64
dynamic inconsistency 527–529
early Chinese philosophy 113; contextual reasoning in Confucianism 113–114; *Mencius* 115–119; practical reasoning and Zhuangzi's intuitive activity 119–123; reasoning from concrete experience 114–115
embeddedness 490–492
emotional reasoning 254–256
emotions 251–260; and moral judgments 277–279
endorsement 189–191
ends 133–134, 134–135
Enoch, David xii, 13–14
epistemology 27–28; of implicit cognition 287

equilibrium: rationally stable reflective equilibrium 522
equivocations 343–344
evaluative propositions 252, 430–431, 434
externalism *see* motivational externalism; motivational judgment externalism
favouring relation 215–216
feeling 117–118
formally valid reasoning 219–220
geometrical model 127–131
GG *see* Guise of the Good Thesis, The
Gibbard, Allan 298, 310, 367, 373, 389
going wrong 505–506
gratuitous costs 509–511
Greenspan, Patricia xii, 11, 252–255, 393–394, 396–397, 401
Guise of the Good Thesis, The (GG) 225–234; the nature of 229–230
Harman, Elizabeth xiii, 15, 543, 561
Haslanger, Sally xiii, 9
hidden mechanisms 285–287
Hieronymi, Pamela xiii, 14
higher order reasons: and emotions 258–260
high-level substantive questions 4
historical reading 554–557
holism 214–215
human agency, theory of 160–161, 170; Bratman and Korsgaard 162–164; Kant 165–169; Leibniz and Clarke 161–162; method 164–165
Hume 60–64, 141, 153; the activity of practical deliberation 146–153; elements of a theory of practical reason 142–143; standards for practical deliberation 143–146
identification 191–192
identity theory 43–47
immersive deliberation 136–138
implicit cognition 287
incommensurability 395, 398, 506, 509–511, 512n16
Independence 527–529
instrumental rationality 482–483; alienation 488–489; explanation of canonical examples 487–488; face-value understanding of 483–486; Necessary Glue Claim 494–495; non-instrumental structure of practical reason 495–498; non-instrumental unity of reason 498–500; preliminaries and disclaimers 483–487; and the Status Explanation Claim 487–500; subsumption and embeddedness 490–492; technical knowledge 489–490; value of rationality 492–493
intentional action 173–174
internalism: conditional vs unconditional 239–241, 246–247; de dicto vs de re 242–243; direct vs deferred 241–242; *see also* motivational internalism; motivational judgment internalism

intuitive activity 119–123
irrationality: metaphysical limits on 200–201; *see also* practical irrationality
judgment: in the *Mencius* 118–119
Kant, Immanuel 47–49, 54–58; constitutivist supplements 326–328; constructing respect for persons 322–326; constructing universalizability requirements 321–322; constructivism 318–333; human agency 160–162, 164–170; motivations for normative constructivism 318–320; reasoning and other people 328–333; varieties of normative constructivism 320–321
Korsgaard, Christine 162–164
Leibniz, Gottfried Wilhelm 161–162
Little, Margaret O. xiii, 14–15
Lord, Errol xiii, 15–16, 46, 407–408
Macnamara, Coleen xiii, 14–15
Markovits, Julia xiii, 13, 313
Mencius 117–118; analogical reasoning 115–117; judgment and values 118–119
metaethics 3–7, 368–370
metaphysical questions 4
metaphysics 26; metaphysical limits on irrationality 200–201; metaphysical propositions about *ought* 430–431
method 164–165; Kant's critical method 165–166
methodology 69–72, 345
moral competence 267–269
morality 441–443, 443–444; authority of 463–464; and subjectivism 314–315
moral judgments: and emotions 277–278, 278–279
morally responsible agency 267–269
moral *ought* 438–441, 446–447, 450–451, 453n1, 455n17–18
moral perspective 444–445
moral psychology, philosophy of 2–3
moral rationalism 264–266
moral requirements: explaining 409–411; and moral *ought* statements 439–441
moral sentimentalism 264–266
moral wrongs 87–90
Morton, Jennifer xiii, 17
motivating reasons, theory of 215
motivation 26–27, 69–72
motivational externalism 237–238, 242–243, 246–248, 248n1, 248n5; *see also* motivational judgment externalism
motivational internalism 237–239; conditional 239–241; de dicto vs de re 242–243; direct vs deferred 241–242; as an empirical thesis 243–246; unconditional 239–241, 246–248; *see also* motivational judgment internalism
motivational judgment externalism 266–267
motivational judgment internalism 237, 266–267

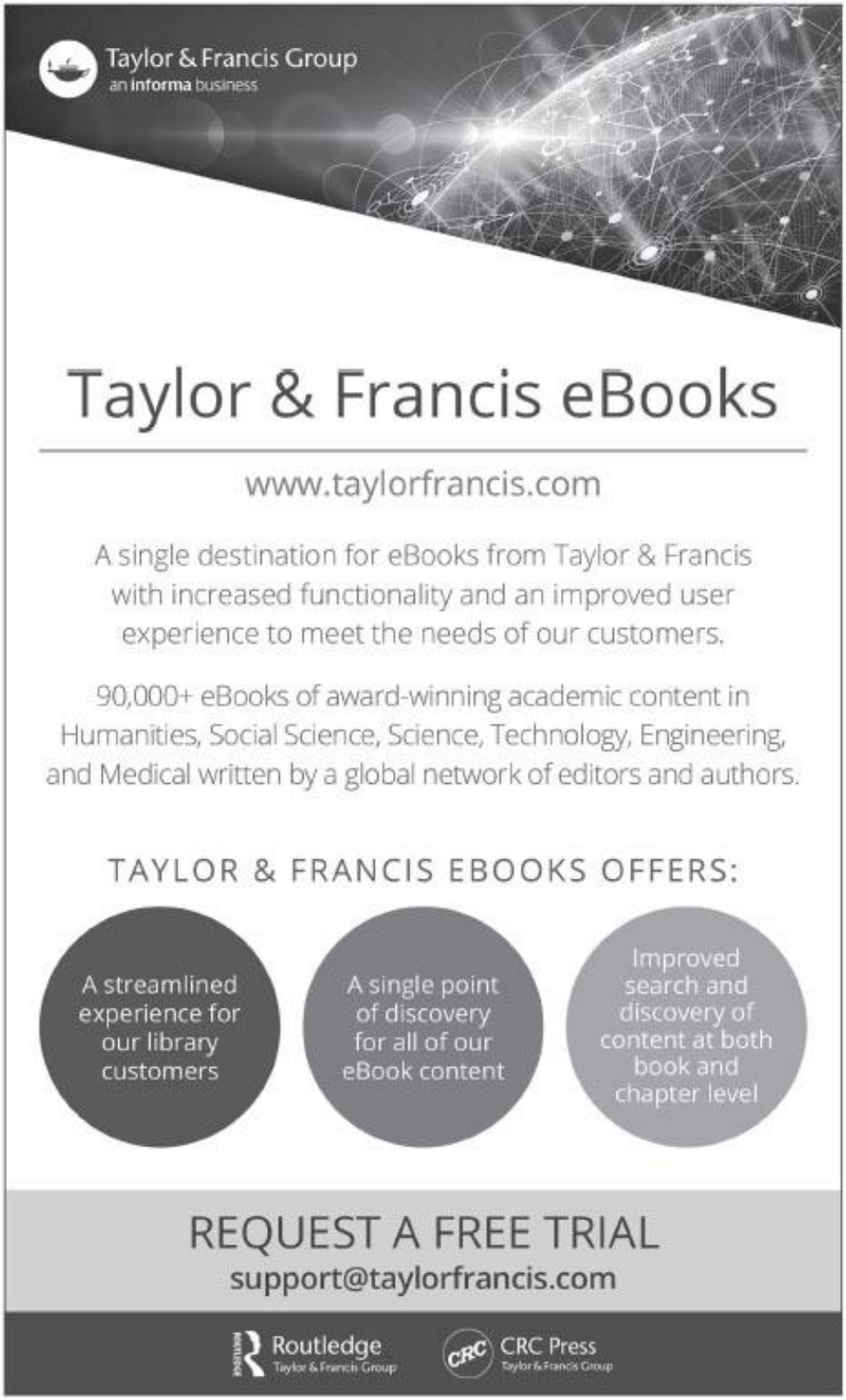
motivations 318–320
narrow-scope coherence requirements 472–473
naturalism *see* normative naturalism
non-ideal theory 9, 69–72
non-instrumental: structure of practical reason 495–498; unity of reason 498–500
nonnaturalism *see* normative nonnaturalism
non-normative truths 34–35
non-requiring reasons 393–402
normative constructivism 318–320; varieties of 320–321
normative force 471–472; alternative views of 476–478
normative impotence: of coherence requirements 472–473; of narrow-scope coherence requirements 472–473
normative institutions: and pragmatism 83–85
Normative Naturalism 376–377; the co-extensiveness argument 377–378; the normativity objection 378–380; scientific analogies 380–386; the triviality objection 386–388
Normative Nonnaturalism 366–368; concepts and properties 373–376; and Gibbard 389; meta-ethics 368–370; against Normative Naturalism 376–388; ontology 370–373; and Railton 388; *see also* normative naturalism
normative pluralism 416–417; an argument for the pluralist view 422–424; and conflict situations 420–421; "metaphysical" and "evaluative" propositions about *ought* 430–431; objections 424–425; objections to the strong pluralist view 431–434; oughts *simpliciter* and the strong pluralist view 425–429; and the unified view 417–420
normative realism 25–28
normative reasons 471–472; sources of 97–98; source voluntarism about 99–100; upshots for 101–103
normativity 38, 41–42, 44–50, 378–380; alternative views of normative force 476–478; alternative views of rationality 478–479; and coherence 470–471; does normativity supervene on the mind 46; does what you ought supervene on the mind 44–46; impotence of coherence requirements 472–473; missing reasons argument 475–476; normative force and normative reasons 471–472; of practical reasoning 544–546; and rationality 43–44, 469–479; of requirements of (structural) rationality 30–31; symmetry argument 473–475
norms: as categorical rational requirements 542–544; contingency and context 546–551; of practical reasoning 541–551; skepticism about 544–546
Objectivism 295–296, 299–300, 302
ontology 370–373
oughts 449–450; metaphysical and evaluative propositions about 430–431; *see also* moral *ought*; prudential *ought*; ought *simpliciter*
ought *simpliciter* 416–417, 419–421, 430–434; and the strong pluralist view 425–429
Parfit, Derek xiii, 5, 12–14, 295–296, 388; instrumental rationality 492, 497; Kantian constructivism 324–325; subjectivism 310–314
passive agency 201–206; desirability and necessity of 207–208

Paul, Sarah xiii, 17
plan rationality 514–517, 520–521; and the strategy of self-governance 517–519, 521–522; and temptation 519–520
pluralism *see* normative pluralism
practical deliberation: Hume on the activity of 149–153; Hume on the capacities required for 146–149; Hume on the standards for 143–146
practical irrationality 197–200
practical rationality 489–490
practical reason 52, 64–65, 262, 270–272; the activity of practical deliberation 146–153; and Aristotle 58–60, 216–217; central questions 25–35; common objections to normative realism 25–28; in early Chinese philosophy 113, 117–123; elements of a theory of 142–143; and emotions 251–260; and the epistemology of implicit cognition 287; good reasoning and overcoming bias 285–287; Hume's theory of 141–153; and Kant 47–49, 54–58; narrow and broad conceptions of 53–54; non-instrumental structure of 495–498; normative truths and non-normative truths 34–35; normativity 38, 41–44, 49–50; norms of 541–551; in philosophical discourse 264–269; philosophy of 1–8; rationalism and sentimentalism about moral judgments 276–280; rationality 38–41, 43–44, 49–50; reasons, action, and agency 281–284; reasons and rationality 29–34; responses to the identity theory 44–47; and the second-person standpoint 457–465; skepticism about the normativity of 544–546; and social practices 68–78, 77–79; and social science research 276–288; the standards for practical deliberation 143–146; and Zhuangzi's intuitive activity 119–123; *see also* practical reasons
practical reason, philosophy of 1–2; defined 2–8
practical reasons 296–298
pragmatism 83–91; and the limits of normative institutions 83–85; and problem-solving 85–87; updating the pragmatist research program 90–91
properties (term) 374–376
prudence 447–449
prudential *ought* 438–439, 447–451
psychological realism 65
psychopathy 262–263, 270–271; in philosophical discourse 264–269
Rabinowicz, Wlodek xiii, 16–17
Railton, Peter xiii, 9, 12, 311–312, 388, 488
rational agency 95–96; the activist view 106; a grounding framework for reasons 96–97; the problem of well-formed choice situations 103–106; sources of normative reasons 97–98; source voluntarism about normative reasons 99–100; upshots for 101–103; and the will 100–101
rational competence 267–269
rationalism 279–281; and moral judgments 276–281; *see also* moral rationalism
rationality 29–34, 38–41, 46–50, 469–470, 505–507, 509–510; alternative views of 478–479; alternative views of normative force 476–478; and coherence 470–471; does rationality supervene on the mind 46–47; normative force and normative reasons

471–472; normative impotence of coherence requirements 472–473; and normativity 42–44; normativity of requirements of (structural) rationality 30–31; the symmetry argument 473–476; theory of 7–8; three ideas of 29–30; value of 492–493; *see also* instrumental rationality; plan rationality
realism 215; *see also* normative realism; psychological realism
reason 52, 65–66; and Aristotle 58–60; and the conditions of action as limits on the possibility of practical irrationality 197–200; and Hume 60–64; and Kant 54–58; and the metaphysical limits on irrationality 200–201; narrow and broad conceptions of practical reason 53–60; non-instrumental unity of 498–500; and passive agency 201–208; and psychological realism 65; and the will 100–101, 196–208; *see also* practical reason; requirements of reason
reasoning 117–118, 349, 360–361; and actions done for reasons 351–354, 357–359; alternative account of reasons 355–356; and bias 285–287; from concrete experience 114–115; current thinking 350; explaining 219; formally valid 219–220; and instrumental rationality 483–486; the "normative" relation 350–351, 359–360; and other people 328–333; the role played by 279–281; vicious people 135–136; and the wrong kind of reason problem 354–357; emotional reasoning; theoretical reasoning
reasons 29–34, 174–177, 281–284; and the agony argument 300–306; commendatory 398–400; as constituents of action 177–179; deontic second-personal 460–463; discounting 256–258; explaining 219; explaining actions done for reasons 357–359; missing reasons argument 475–476; objectivism about 295–306; permissibility-conferring 395–398; second-personal 459–460; self-regarding 444–445; subjective theories of 298–300; theory of 214–215; two kinds of theory of practical reasons 296–298; the wrong kind of 354–357; *see also* acting for reasons; higher order reasons; non-requiring reasons; normative reasons; practical reasons
reasons, theory of 214–215; *see also* motivating reasons, theory of
reflection 117–118
reflective deliberation 136–138
regret 505–511
relational account 411–413
requirements of reason 405–414; basic idea of 405–407; explaining 407–409; *see also* moral requirements
resolution/resolute choice 527, 531–537
respect for persons 322–326
Rosati, Connie xiii, 11
Sayre-McCord, Geoffrey xiii, 10
Scanlon, T. M. xiii, 5–9, 12, 15, 268, 350–352; and *oughts* 418; and subjectivism 310–311
Schapiro, Tamar xiii, 10, 76–77
scientific analogies 380–386
second-person standpoint 457–465
self-governance 517–519, 521–522
self-regarding reasons 444–445

sentimentalism 276–281; *see also* moral sentimentalism
Shmagency 339–342
shuffling 505–506, 510–511, 519
Singh, Keshav xiii, 10
skepticism 544–546; *see also* normative skepticism
slavery 87–90
Smith's constitutivism 337–339
Sobel, David xiii, 13
social science research 276, 287–288; good reasoning and overcoming bias 285–287; rationalism and sentimentalism about moral judgments 276–281; reasons, action, and agency 281–284
social practices 68–69, 74–77; and culture 72–74; methodology, motivation, and non-ideal theory 69–72; relevance of 77–79
sophistication/sophisticated choice 527, 529–530; disadvantages of 530–531; and resolute choice 531–532; and wise choice 533–537
source voluntarism 99–100, 107n10
subjective theories 298–306
subjectivism 307–316
(structural) rationality 30–31
Sylvan, Kurt xiv, 16, 362n32, 493, 498
technical knowledge 489–490
temporary preference reversals 506–507
temptation 506–507, 519–520
Tenenbaum, Sergio xiv, 11
theoretical reasoning 216–217
theory *see* agency, theory of; human agency, theory of; identity theory; motivating reasons, theory of; non-ideal theory; rationality, theory of; reasons, theory of; *and under* practical reason
Tiberius, Valerie xiv, 12
triviality objection 386–388
truths 34–35
unity of reason 498–500
universalizability 321–322
vague goals 507–509
value of rationality 492–493
values 118–119
valuing 307–309
vice 133–136
voluntary 356–357
Walden, Kenneth xiv, 12–13
Wallace, R. Jay xiv, 14–15, 43, 496, 543
warrant for emotion 253–254

Washington, Natalia xiv, 12
well-formed choice situations 96, 103–106
will, the/willing 99, 196–197; and the conditions of action as limits on the possibility of practical irrationality 197–200; creating reasons 100–101, 106; and the metaphysical limits on irrationality 200–201; and passive agency 201–208
wise choice 527, 533–537
Wonderly, Monique xiv, 11–12
Wong, David xiv, 9
wrong kind of reason problem 354–357
Zhuangzi *119–123*


Taylor & Francis Group
an informa business
Taylor & Francis eBooks
www.taylorfrancis.com
A single destination for eBooks from Taylor & Francis with increased functionality and an improved user experience to meet the needs of our customers.
90,000+ eBooks of award-winning academic content in Humanities, Social Science, Science, Technology, Engineering, and Medical written by a global network of editors and authors.
TAYLOR & FRANCIS EBOOKS OFFERS:
A streamlined experience for our library customers
A single point of discovery for all of our eBook content
Improved search and discovery of content at both book and chapter level
REQUEST A FREE TRIAL
support@taylorfrancis.com
Routledge
Taylor & Francis Group
CRC Press
Taylor & Francis Group

The Ted K Archive

Kurt Sylvan & Ruth Chang
The Routledge Handbook Of Practical Reason
Routledge Handbooks in Philosophy
2020

<www.routledge.com/The-Routledge-Handbook-of-Practical-Reason/Chang-Sylvan/p/book/9780367695408>
ISBN 0429266766, 9780429266768

Routledge

www.thetedkarchive.com