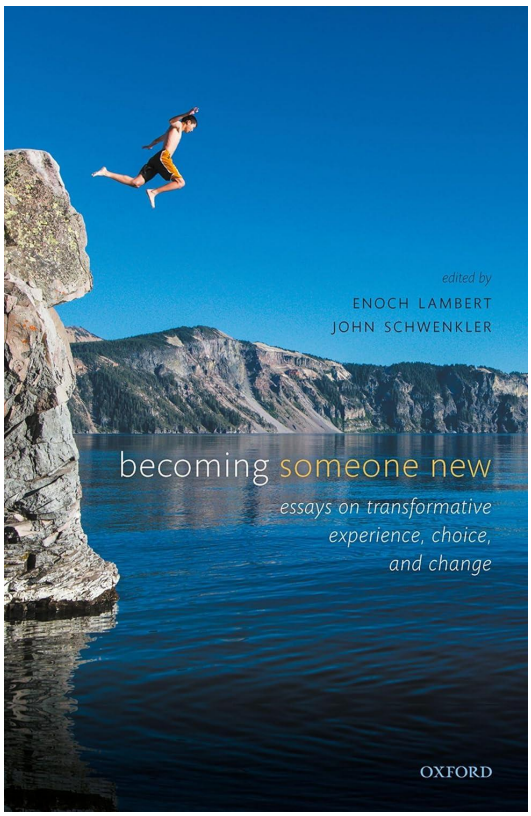
Becoming Someone New

Essays on Transformative Experience, Choice, and Change

John Schwenkler & Enoch Lambert



2020

Contents

Becoming Someone New
Essays on Transformative Experience, Choice, and Change
EDITED BY
Enoch Lambert
and John Schwenkler
OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS
Great Clarendon Street, Oxford, 0x2 *6dp,*
United Kingdom
Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship, and education by publishing worldwide. Oxford is a registered trade mark of Oxford University Press in the UK and in certain other countries
© the several contributors 2020
The moral rights of the authors have been asserted
First Edition published in 2020
Impression: 1
All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, without the prior permission in writing of Oxford University Press, or as expressly permitted by law, by licence or under terms agreed with the appropriate reprographics rights organization. Enquiries concerning reproduction outside the scope of the above should be sent to the Rights Department, Oxford University Press, at the address above
You must not circulate this work in any other form
and you must impose this same condition on any acquirer
Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America
British Library Cataloguing in Publication Data
Data available
Library of Congress Control Number: 2019954845
ISBN 978-0-19-882373-5
Printed and bound in Great Britain by
Clays Ltd, Elcograf S.p.A.

Links to third party websites are provided by Oxford in good faith and for information only. Oxford disclaims any responsibility for the materials contained in any third party website referenced in this work.

List of Contributors

Nomy Arpaly is Professor of Philosophy at Brown University.

Katalin Balog is Professor of Philosophy at Rutgers University, Newark.

Daniel M. Bartels is Professor of Marketing in the Booth School of Business at the University of Chicago.

Agnes Callard is Associate Professor of Philosophy at the University of Chicago. *Matthew Cashman* is a graduate student in Cognitive Science at the Massachusetts Institute of Technology.

Stephanie Y. Chen is Assistant Professor of Marketing at London Business School. *Fiery Cushman* is John L. Loeb Associate Professor of the Social Sciences in the Department of Psychology at Harvard University.

Martin Glazier is Wissenschaftlicher Mitarbeiter in Philosophy at the University of Hamburg.

Amy Kind is Russell K. Pitzer Professor of Philosophy at Claremont McKenna College. *Jennifer Lackey* is Wayne and Elizabeth Jones Professor of Philosophy at Northwestern University.

Enoch Lambert is a Postdoctoral Associate in the Center for Cognitive Studies at Tufts University.

Sarah Molouki holds a Ph.D. from the Booth School of Business at the University of Chicago.

L. A. Paul is Professor of Philosophy and Cognitive Science at Yale University. *Richard Pettigrew* is Professor of Philosophy at the University of Bristol. *Nick Riggle* is Assistant Professor of Philosophy at the University of San Diego. *John Schwenkler* is Associate Professor of Philosophy at Florida State University. *Rosa Terlazzo* is Associate Professor of Philosophy at the University of Rochester. *Evan Thompson* is Professor of Philosophy at the University of British Columbia. *Tomer Ullman* is Assistant Professor of Psychology at Harvard University.

Oleg Urminsky is Professor of Marketing in the Booth School of Business at the University of Chicago.

Samuel Zimmerman is cofounder of Freebird.

Editors' Introduction[(1)]

Enoch Lambert and John Schwenkler

1. Cleo Considers Transformation

Imagine Cleo, who is pondering taking a continuing education course at Local Org, a local institution that primarily serves the elderly. Cleo wasn't able to afford much college when she was young, and wishes to take advantage of learning opportunities in retirement. She has already attended a couple of courses on topics like art history and identifying local bird species, which she enjoyed. Now she is considering one that has been causing a stir among her friends and acquaintances at Local Org. The course, titled "Radical Perspectives on Economics: From the Household to Global Trade," has divided those who have taken it so far. Some say the ideas taught are pernicious and simply intended to stir people up. A few are so mad they are quitting their membership with Local Org, and are even suggesting lobbying the city council to discontinue its funding. Others describe having their view of the world profoundly altered. Some feel invigorated and emboldened as a result, changing their priorities to activist-oriented ones. And some describe what they've learned as causing profound rifts with their spouses and other family members as they retrospectively re-evaluate the distribution of burdens in their relationships. A few feel as though their whole understanding of their lives in society has been based on lies and are still trying to figure out the consequences. There are more takes besides, but virtually everyone who has taken the course appears to have been profoundly moved.

Cleo is intrigued. She's very curious and wants to find out for herself what all the fuss is about. On the other hand, she was looking forward to a fairly calm retirement, engaged in projects and activities she had been planning for years, even decades, and is concerned that taking this course may profoundly disrupt her future. Of course, she reasons, doubtless there are things that should disrupt her in these ways, including things revealed by courses such as this one. Cleo really wants to take the course but also has the nagging feeling that she just doesn't know how she will respond to it. But then, arguably the only way to figure that out is by giving the course a try.

[(1)] Enoch Lambert and John Schwenkler, Editors' Introduction In: Becoming Someone New: Essays on Transformative Experience, Choice, and Change. Edited by: Enoch Lambert and John Schwenkler, Oxford University Press (2020). © Enoch Lambert and John Schwenkler.
DOI: 10.1093/oso/9780198823735.003.0001

Cases like these raise questions. There is, of course, the immediate practical question that Cleo herself faces: What should she do—take the course or not? Let us assume for the sake of argument that she should choose whichever path she anticipates being more satisfying or fulfilling. The trouble is that she's been given reason to believe that taking the course can change what she finds satisfying, and in significant ways. It may cause her to reevaluate how fulfilling her life has been thus far, as well as her plans for her retirement years. Thus the immediate practical question gives rise to a number of further ones: How should Cleo evaluate and compare the salient possible outcomes she sees around her? Where can she find the critical information she needs to make the decision? What should she make of the fact that taking the course changed her friends in ways that they themselves did not anticipate? Changes it seems they couldn't have anticipated, for they partly consist in responses to knowledge available only through the course of the change itself. What will Cleo make of the life she has led—her struggles, joys, efforts? How will her selfunderstanding be altered?

The forward-looking part of Cleo's deliberation provokes a related set of questions. Her current preferences and desires lay out a certain vision for her retirement years. And parts of that vision are incompatible with the activist-oriented lifestyles that some of the course's previous participants are now pursuing. She can't well imagine being moved by the passions now moving them. And the way they now talk comes across as alien to her way of thinking. Indeed, Cleo has a hard time imagining how a course could inspire any of the range of inflamed reactions she's been witnessing. So the outcome of taking this course is somewhat opaque to Cleo. Still, she is interested in learning some economics, is intrigued by the course description and materials, and desires to join the conversations that the course has been sparking. One difficulty that Cleo faces in her decision-making is that she doesn't seem able to adequately comprehend the outcomes in advance. And another is that several of the outcomes to be considered involve significant personal change. They involve changes that are often described as transformative.

Acknowledging the phenomenon of transformative change leads in turn to a number of philosophical questions that have to do with the very possibility of making life-altering decisions in a rational way. For decisions with lesser stakes (do I give up weekly basketball or movie night to finish this work presentation?) we apply our usual, familiar standards, letting ourselves be guided by our standing desires, preferences, values, and reasons for action. But what happens when the outcomes to be evaluated promise to disrupt these very standards? On what basis can we judge that which threatens to transform our basis of judgment? If our desire for first-hand experience rises with the stakes of a decision, then it seems like statistics and testimonials will not suffice. But the only way to gain the desired familiarity would be to make the decision in favor of transformation.

Finally, other philosophical questions arise concerning the very nature of transformative change of the sort that Cleo has to choose or reject. What kind of change revamps a person's whole perspective? What are the psychological mechanisms by

which it operates? What is it like to experience such a change? Is it possible to imagine this experience accurately in advance? If we could, would that be enough to enable a rational decision about whether to undergo such a change? And what is the metaphysical relation between a person as she exists prior to a transformative change and the person who exists in the wake of it? That is, is it a relation of Lockean personal identity or non-identity, Parfitian continuity or discontinuity, and so on? As with the first set of philosophical questions raised in the previous paragraph, the way we resolve these things seems to have a direct bearing on the practical questions that a person facing a potentially transformative change will want a way to settle.

The kinds of choice-induced changes Cleo considers are termed transformative by Edna Ullmann-Margalit (2006) and L. A. Paul (2014, 2015c). Paul's writings in particular have generated a great deal of interest in and work on the topics of transformative choice, change, and experience across several different disciplines. This volume contributes 15 new essays to this ongoing conversation. There have already been a number of direct responses to Paul's arguments, many of which are collected in journal volumes along with Paul's own clarifying defenses and extensions (see *Res Philosophica* 92(2); *Philosophy and Phenomenological Research* 91(3)),[1] and some of those exchanges are continued here. But all of these essays also break some new ground, either employing notions of the transformative in novel philosophical or scientific contexts or bringing established ideas into dialogue with them. There are brief summaries of each chapter at the end of this introduction.

We regard these essays as exemplary contributions to continuing research, and trust they will prompt more work on these exciting topics. In that spirit, we focus the rest of this introduction on outlining some issues that we think deserve further engagement. We begin by returning to Ullmann-Margalit's work for insight. If we are to take the first-personal perspective on transformative change as seriously as Paul recommends, a good way to proceed is to describe differences in the ways people experience it, as Ullmann-Margalit does. Turning then to Paul's work, we draw out some of her distinctive contributions while also critically examining some of her presuppositions. Finally, we note that Paul and Ullmann-Margalit both share what we call a *Replacement Model* of what transformative change amounts to, and conclude by raising some questions about this assumption.

2. Ullmann-Margalit on Types of Decisions

Transformative changes are irreversible changes to the "core" of who one is (Ullmann-Margalit 158; cf. Paul 2014: 16). In her work, Ullmann-Margalit employs the labels "Old

[1] Paul's extensive work communicating these ideas beyond the academy is collected at her website: <`https://lapaul.org/news.html`>. She also has a helpful teaching guide, with summaries of central ideas as well as reading lists: <`https://lapaul.org/papers/teaching-guide-for-transformative-experience.pdf`>.

Person" and "New Person" to designate individuals pre- and post-choice, and "transformation" as the term for the change from one to the other (2006: 167). Putting things this way helps capture some important aspects of the puzzles that arise in connection with different cases. If self-interest partially makes an individual who they are, then how can it be in the self-interest of an individual to repudiate their own present interests, and take on new ones instead? If an individual's core values partially determine who they are, then what could possibly determine which set of wholly different values to adopt? More generally, whatever one uses as the basis for all their choices, what about *that basis* could make it rational to choose a shift to a new basis?

An important feature of Ullmann-Margalit's and Paul's work is their concern to motivate their philosophical puzzles both intuitively as well as in "technical" terms. That is, each argues that the phenomenon of transformative change challenges some aspect of our common sense, pre-philosophical approaches to rational decisionmaking, while also calling into question standard formal decision models of what rational choice involves. At an intuitive level, what many people find adverse in standard optimization-pursuing strategies is the presumption that big decisions can be evaluated with calculative balance sheets (Ullmann-Margalit 2006:165). Somehow such an approach is supposed to be out of step with the significance of such decisions, which can really only be made via some more intuitive mechanism. One way to understand Ullmann-Margalit's project is as a way of vindicating this unease by showing that it is keyed into a genuine limitation of standard decision theories.

To do this, Ullmann-Margalit employs the term "rationality base" to describe those psychological features of a person in virtue of which they are able to make rational decisions.[2] She further distinguishes decisions big, medium, and small, where medium ones are those most clearly subject to standards of rationality. For Ullmann- Margalit, a small decision, of the sort one faces in wanting a Coke while in front of a wall of shelves full of them, doesn't present a true choice. One doesn't choose which Coke to drink, but merely "picks," as one is indifferent between the options (2006:157). By contrast, a medium choice would be one over whether to buy the tempting Coke at all, when one is trying to build healthier diet habits. Finally, big decisions are the ones that promise to bring about transformations in one's rationality base itself—for example, a corporate executive considering whether to change course and pursue a life of Buddhist renunciation (2006:160). Ullmann-Margalit calls the move to change course in such a way "opting," and asks whether optings can be evaluated rationally, like medium decisions, or if they are more akin to "picking," as in the case of selecting a Coke.

About such big decisions, Ullmann-Margalit writes:

> Opting transforms the sets of one's core beliefs and desires. A significant personality shift takes place in our opter, a shift that alters his cognitive as

[2] Ullmann-Margalit talks only about such bases consisting of beliefs and desires (or preferences

> well as evaluative systems ... The question I am raising is whether it is possible to assess the rationality of his choice, given that this choice straddles two discontinuous personalities with two different rationality bases. (2006: 167)

There are two things at issue here. One is the overall psychological discontinuity between Old Person, the deliberating subject who must choose for or against a transformative change, and New Person, the subject whom this transformative change would generate. The second is one particular aspect of that discontinuity, namely the inconsistency between the rationality bases that characterize these two respective personalities. New Person's new sets of beliefs and desires may well be "internally consistent," but the point about the transformation is that "inconsistency" now exists between New Person's system of beliefs and desires, taken as a whole, and Old Person's system taken as a whole (Ullmann-Margalit 2006: 167).

Ullmann-Margalit goes on to argue that for both picking and opting, the standard rational strategy of *optimization* doesn't make sense (2006:168). In cases that require picking, this is due to subjective indifference. There's no rational strategy to adopt, other than that of not wasting any time thinking about it. With opting, by contrast, there is no such indifference: it *matters* to us that we make the best decision we can in these situations, and we have a strong sense that some choices could be better or worse than others, even as it is hard for us to say what this goodness and badness consists in.

As we indicated above, standard decision theories pursue optimization, e.g. *maximization* of expected utility (Von Neumann and Morgenstern 1953) as a basis of rational choice. And any real optimization problem needs constraints to be tractable. In the case of rational choice, we need constraints on the "rationality base" (beliefs and desires, credences and preferences, etc.) of the individual who is choosing. In standard decision theories, such constraints are largely internal to the rationality base, i.e. constraints on how the base is structured. One example often used to illustrate this is transitivity in preference structure: a rational preference ordering will not allow preferring C to A when A is preferred to B and B is preferred to C.

Ullmann-Margalit glosses such constraints as a requirement that rational bases be "internally consistent." Further, she argues, the problem with big, transformative decisions is that they require leaps from one rationality base to another. Even where each base is internally consistent they may, as we have seen, be inconsistent *with one another*. Ullmann-Margalit does not elaborate on the extent to which such consistency can be put in formal terms. Intuitively, though, her examples illustrate cases where core desires of people pre- and post- choice would be incompatible with each other (e.g. the ambition of a CEO finds no place in Buddhist renunciation). If a rationality base determines a specific optimization problem (i.e. how to maximize *this* person's

when focused on formal decision theory), though theories of what enables rationality may differ over

expected utility), there seems to be something wrong, perhaps even nonsensical, about tasking it with "choosing" to become a person with a different rationality base with a different optimization problem, as though the choice itself were just another run-of-the-mill attempt to optimize. The usual goals and constraints on optimizing don't seem to do the trick here, and Ullmann-Margalit thinks that some people's intuitive resistance to calculative rationality in big decisions is a reflection of this fact.

But if we cannot make big decisions rationally in the standard way, Ullmann- Margalit thinks we might still approach them *reasonably* by conforming to standards other than the usual optimizing ones: "'Acting rationally' need not mean optimizing; it can also mean acting reasonably" (2006: 168). She briefly sketches a few candidate standards of reasonability, where one consists in being clear with oneself about the weight of big decisions and the fact that one has to make them anyway. The alternative is a kind of self-deception where one simply "drifts" into one option or another, a course of behavior which is less reasonable by Ullmann-Margalit's lights. A superior possibility is to work on getting clearer about one's higher-order preferences to see if these can help light the way. For surely, we might think, a person's higher-order preferences will be sufficient to rule out *some* first-order preference structures. Ullmann-Margalit considers this but does not pursue it in detail. She comes across as ambivalent about its prospects (2006: 168). And a third route to reasonability is through what might be called approximating choice, i.e. using trial runs or partial tests, where available, to turn our big decisions into something more like middle-sized ones. Ullmann-Margalit offers the practice of living with someone prior to marriage as an example (2006: 169).

Ullmann-Margalit's attention to alternative means and standards of decision making parallels the satisficing and bounded rationality research traditions as alternatives to standard decision theory (e.g. Simon 1956; Gigerenzer and Selten 2002). Determining the extent to which insights from these and similar traditions might inform research on transformative choice, and visa versa, is work that remains to be carried out. Although she does not offer final resolutions to all the puzzles she raises, some of the value of Ullmann-Margalit's work lies in the way it reminds us that there are multiple kinds of decision contexts where standard decision theory either distorts the way we actually make decisions or fails to offer proper guidance concerning them.

3. Paul and Subjective Experience

In Paul's work, the principal ways in which transformative changes are taken to raise philosophical questions are through their subjective effects—effects which she argues cannot be properly anticipated or imagined by an individual who must choose for or against them, or grasped adequately except by those who have been through the changes themselves. Paul makes a distinction between transformative change that is

what makes up people's rationality bases, how they are structured, and so on.

epistemic and transformative change that is *personal* (Paul 2014: 10-16). The former is induced via experiences with novel phenomenal features (e.g. novel food or the experience of a colored world, as in Frank Jackson's (1986) famous case of a colorblind neuroscientist). The latter results in changes to one's personal "core." Paul focuses in particular on changes to our core *conative* characteristics and how they affect our experience of ourselves and our world. And her central thought, illustrated well by the example of deciding whether or not to have a child, is that novel experience and consequent changes to one's subjectivity make it impossible to assign the subjective values necessary to *compare* the prospect of having a child with that of going forward as one so far has (Paul 2015c).

A helpful way to conceive of Paul's attack on standard decision theory is as a challenge to the so-called *completeness axiom*, fulfilment of which is supposed to be necessary for constructing utility functions out of preferences. Simplifying somewhat for present purposes, this axiom says that for any pair of choice alternatives, an individual necessarily prefers one or the other, or is indifferent between them. It has been recognized by many, starting with its authors in the first formal decision theory, to be the most intuitively unrealistic of the axioms.[3] Paul's achievement is to turn this intuition, of which there are scatterings of sometimes unrealistic or extreme examples in the literature, into a seemingly robust class of clear and common choices in which the requirement of completeness cannot be satisfied. If Paul is right, the phenomenon of transformative choice reveals far-reaching shortcomings in standard approaches to rational decision-making.

On one natural reading, at the core of Paul's argument is the claim that there is a class of possible choices concerning which we don't have any preferences formed in advance. This reading is plausible, but with the caveat that at least some of what Paul would want to count as personally transformative experiences are such that people very often do have clear preferences, both stated and revealed, in favor of some possibilities over others. Many people have explicit preferences for and against having children, for example, and Paul's arguments are aimed at them as well. Of course, for other transformative choice situations there may be no clear preference (think of situations like those faced by Cleo, where one's advance preferences seem hard to articulate). Future work should distinguish transformative choices where people have clearly stated preferences from those where they don't, and should consider whether decision theories need to treat them differently.

We should, then, supplement the claim above with the further observation that in situations of potentially transformative choice, not all of our preferences—stated, revealed, or otherwise—are adequately *justified* or in proper normative standing. That is, preferences relative to possible personally transformative experiences need to be

[3] "The axiom ... expresses ... the completeness of the individual's system of preferences. It is very dubious, whether the idealization of reality which treats this postulate as a valid one, is appropriate or even convenient" (Von Neumann and Morgenstern 1953, 630). See also Aumann 1962.

informed in the right way in order to determine *rational* choice on the basis of them. But, Paul observes, many such preferences can only be properly informed via subjective acquaintance with the results of choosing one way rather than another. And her argument is that this necessary acquaintance is simply impossible, or at least that it is not yet understood how we could obtain it.

On this understanding of Paul's argument, there will be problems in formulating general principles which suffice to capture the relevant norm. Against the necessity of acquaintance, our biological constitution and human needs make it impossible to think that *all* our preferences should be backed by some sort of direct acquaintance. Against the sufficiency of acquaintance itself, the dynamics of preference change, prevalence of self-ignorance, practices of trying things out, context-sensitivity of tastes, and so on all make any simple appeals to acquaintance as a proper ground of preference formation hopelessly naive.[4] So there remains work to do here as well.

4. Assimilation Strategies

Paul's argument, like Ullmann-Margalit's, is directed at forms of standard decision theory that conceive subjective rationality as a matter of weighing the utility of decision alternatives (assigned via the completeness axiom) with probabilities that the alternatives will occur. The assignment of probabilities is governed by different principles along with one's available evidence, which for many reasons cannot plausibly be restricted to what one has had direct acquaintance with. Probability distributions are made to deal with subjective uncertainty. If one is insufficiently certain about possible assignments of utility, one might plausibly think that you can just extend the probability distribution to cover such assignments. Jessica Collins (2015) and Richard Pettigrew (2015) have each developed different ways of pursuing this kind of strategy formally. It might, then, be thought that once the technical details are worked out, they will show how a person can apply evidence, in just the way we usually do, in making potentially transformative choices. People might seek scientific studies or testimony from loved ones, trusted experts, etc. to assign probabilities to possible utility outcomes of their personally transformative experiences (see Dougherty etal. 2015 for one defense of such strategies). We may classify responses in this style as part of a family of Assimilation Strategies, as they seek to "assimilate" the treatment of uncertainty about utility to familiar forms.

In previous work, and also in her contribution to the present volume, Paul objects to assimilation strategies on the grounds that they "alienate" a person from their choices. At best, lack of acquaintance with the subjective values that determine which way one should go leaves a person with a certain lack of understanding or inability to identify with the basis for their decisions. At worst, it could put a person in the position of

[4] Vanderbilt 2016 ably surveys scientific research and case studies that undermine even our first personal perspectives on our preferences and desires.

surrendering her deepest desires to "impersonal" data and external advice. Paul asks us to imagine the case of Sally, who deeply yearns to become a mother and believes it will bring her the most happiness and fulfillment (2014: 87-8). Adopting an Assimilation Strategy could lead to Sally deciding to remain childless due to the best scientific data of the day. Paul argues that a world in which we surrender our first personal perspectives in the most important decisions is unacceptable:

> To be forced to give up the first person perspective in order to be rational would mean that we were forced to engage in a form of self-denial in order to be rational agents. We would face a future determined by Big Data or Big Morality rather than by personal deliberation and authentic choice. (Paul 2014: 130)

One way to think about Sally, or someone similar to her, is as aspiring to motherhood. She acknowledges it as something that she understands dimly at best, but wishes to grow into an understanding of by way of direct engagement with it. Human life without aspiration will appear as bizarre and alien to many as anything conjured by Big Data (or Big Morality or Rationality). In fact, one might think that any theory of rationality that makes the pursuit of aspirations either irrational or even non- rational should be taken to be a non-starter. Any theory seeking to capture people's decision-making frameworks needs to take account of how thoroughly they are structured by their aspirations.

Agnes Callard (2018) has developed a framework for aspiration, arguing that its proper treatment requires recognition of what she calls proleptic reasons. She differs from Paul on many issues, including whether our aspirations are properly thought of as things that we make decisions about. We need not endorse Callard's framework to recognize the centrality of aspiration or aspiration-like desires in human life. If aspirational decisions cannot be held to Paulian standards of acquaintance, they might nevertheless indicate philosophically important features of human motivation and decision-making. It may be, for instance, that we are largely unaware of the conflicts that are bound to arise between our current preferences and the necessary changes that attend pursuing our aspirations. More generally, our jumble of desires, values, motivations, and so on may conflict in any number of ways that are either ignored or idealized away by standard decision theory. Closer attention to aspiration may help us construct more realistic models of human motivational structure and decision-making.

5. Models of Transformation

In pressing their respective cases against standard decision theory, Ullmann-Margalit and Paul both employ versions of what we'll call the *Replacement Model* to characterize the type of change that transformations occasion. In personal transformation, a "new" self or personality comes into being, replacing the "old" one who chose

to go in for the transformation in question. Both distinguish the subject of personal transformation from what philosophers have traditionally argued over under the guise of *personal identity*.[5] Ullmann-Margalit (2006: 167) coins the term *personality identity* to designate what transformation replaces, and Paul speaks of *selves* (20i5d: 490). As Paul puts it (2015c: 798), of all the selves that transformative decision-makers may consider as possible future selves, selves resulting from transformative change are not among them. This helps drive home Paul's *alienation objection* to reliance on non-first-personal sources of information to make transformative decisions. The potential alienation is such that it amounts to asking someone to make a decision to *extinguish* herself in the absence of any authentically first-personal basis for this decision.

In addition to the work it does in arguing for the inadequacy of standard decision theory to the phenomena of transformative change, another attractive feature of the Replacement Model is the way it articulates a common experience that philosophers have often had difficulty in characterizing. Many people who have undergone dramatic life changes do sometimes talk about being completely new or different people. This kind of practice is even institutionalized in some religious and other cultural settings, where personal conversion is treated as an important phenomenon. Ullmann-Margalit suggests that conversion experiences are a similar but distinct class from transformative ones, with the primary dimension of difference being the role of choice. She says that many report experiences of conversion as being irresistible—as presenting no "viable alternatives" at all, though the described effects are every bit as dramatic as any transformative experience (2006:161-2). Attempting to understand such experience without being beholden to the quite different concerns of the philosophical literature on personal identity makes salient the need for a distinction like the ones that Ullmann-Margalit and Paul both draw. But the distinction raises philosophical problems of its own. Personal identity, in the analytic philosopher's sense, can also be of obvious concern to our decisions and to our self-interpretations. So, do we have *two* things, personal identity and personality or self-identity, each of which is of ultimate concern in our deliberations?

Whatever the implications of the above, a more immediate concern is whether a Replacement Model really serves the purposes it is designed for. Let's examine the solution Paul proposes toward the end of her book (see Paul 2015a: 105-24). Her Discovery Solution proposes that we can rationally choose for or against transformation by consulting our preferences for discovery. Do we prefer to discover the kind of person we would become by making the transformative choice, or do we prefer not to? If the Replacement Model is correct, however, then the self that chooses transformation would not discover what it is like to become the transformed self. This is because it, along with its capacities and preferences, would cease to be.

[5] There is some precedent for distinguishing traditional "personal identity" from something like a first- personal decision making perspective. See, e.g. Velleman 1996, whom Paul draws a connection to in her 2015b.

Paul seems to be aware of the issue just raised, or something close to it. In the chapter developing the Discovery Solution, she takes care to emphasize both that we cannot know what transformed experience will be like and that we cannot know what core (i.e. self-defining or -determining) preferences transformation will bestow (see Paul 2015 a: esp. 116-20). We must realize just how much of a leap into the dark must be preferred for the solution to work (2016a: 121-3). And yet, there is no reason to think that even so "prepared" a preference as one in favor of Discovery would be any less subject to the effects of transformation than any other core preference. So far as Paul has given reason to consider, Discovery preferences are as susceptible (or resistant!) to transformation as any others (see e.g. Paul 2015a: 116). In Paul's terms, a transformed self may have no interest whatsoever in discovering the unfolding of its preferences and attendant experience. And crucially, on any Replacement Model, it won't be the self that acted on the discovery-preference that is later on doing the discovering.

Of course, it certainly seems plausible that we can discover the changes we go through, including transformative ones. Many will say they have done it![6] We do not mean to deny this reality. To the contrary, our purpose in making a case for its incompatibility with Replacement Models of transformation is just to argue that additional models of what transformation amounts to, and what it is to undergo it, need to be articulated and explored. In doing so, one thing to keep in mind is that ignorance concerning which core preferences will undergo change does not imply anything about the extent or character of transformative change itself. Imaginative limits are also of limited significance here. A prospective parent may not be able to imagine any novel phenomenology accompanying the love that they will develop for their children, but they can be fairly confident they will develop such love, and that it will be recognizable as love, even though it transforms their understanding of what loving someone can amount to. Data points such as these, acknowledging paths of continuity or development across transformative change, must also be accounted for in models of transformation.

One of our aims here has been to convey how ripe for further inquiry issues surrounding transformative experience remain. Both Ullmann-Margalit and Paul provided groundbreaking conceptual tools and framing for such inquiry. We've suggested places where we think alternative framing deserves exploration. But the proof of the pudding for how rich this topic is lies in the work done by our authors, and so we turn to them.

6. Chapter Summaries

The chapters in this volume have been arranged in a way that is meant to maximize points of overlap in topic and methodology, though there might be many other possible arrangements that would have done this just as well.

[6] Indeed, it's unclear how to understand what is going on with Paul's hypothetical transformed vampires in her book's opening chapter without interpreting them as having discovered the changes they

In "Who Will I Become?" (Chapter 1), L. A. Paul revisits the account of transformative experience developed in her earlier work, discussing the way that lifechanging experiences like having a child or taking on a new career are hard to evaluate within standard models of rational decision-making, given the extent of the change that they bring about in our preferences and the subjective inaccessibility of the phenomena they center on. She argues, further, that testimony from others who have had the experiences in question does not on its own allow a person to overcome this barrier, since in relying on the judgments of others a person alienates herself from her own subjective perspective, reasoning from the point of view of someone that she herself is not. For Paul, a transformative choice requires a person to adjudicate between incommensurable perspectives or self-understandings, and the upshot of transformation is to have one's current self replaced with someone new.

Martin Glazier's "Being Someone Else" (Chapter 2) picks up this last thread, asking how we can make sense of the idea that there is some contingency in the fact that a given person is the person she in fact is. To do this, Glazier develops a way of understanding our ordinary concept of a perspective from which a person experiences and thinks about the world. On the account he offers, the contingency of being who one is arises from the fact that there are perspectives other than one's own from which the world can be apprehended veridically—which gives rise in turn to the possibility of one's own perspective on the world having been one of these. As Glazier notes, however, nothing in this possibility entails that a person is able actually to become someone different, at least in the sense discussed in his chapter. To make sense of this more radical idea, a different framework would be required.

Sarah Molouki, Stephanie Y. Cheng, Oleg Urminsky, and Daniel M. Bartels, in "How Personal Theories of the Self Shape Beliefs About Personal Continuity and Transformative Experience" (Chapter 3), consider how individual beliefs about the personally disruptive character of transformative experience are influenced by intuitive theories of what a self fundamentally is. Their studies focus on two aspects of these intuitive theories. First, they explore how views of the causal centrality of a trait to a person's self-concept influence judgments of whether a change in that trait constitutes a disruption in who that person is. Following this, they build on past work in exploring the extent to which judgments of the personal disruptiveness of a transformative change are influenced by whether the change is perceived as having a positive or negative valence.

Chapter 4, Samuel Zimmerman and Tomer Ullman's "Models of Transformative Decision Making," begins by developing an empirically informed model of the psychological architecture involved in making transformative choices. The aim of this model is to account for the subjective uncertainty emphasized by Paul in her framing of the problem of transformative choice, while showing how an appeal to higher-order pref-

went through *as* changes in their previous selves, and as retaining the capacity to compare "intraself" experiences.

erences can resolve the uncertainty that these choices necessarily present us with. In the second part of this chapter, they employ this model to explore what they call the tyranny of the present in the way that transformative choices are evaluated.

Richard Pettigrew's "Transformative Experience and the Knowledge Norms for Action" (Chapter 5) is another formal treatment of this problem of subjective uncertainty. Having laid out the problem, Pettigrew then discusses an argument due to Sarah Moss that it is impossible for a person facing a transformative choice to know in advance what her post-transformation preferences are likely to be. In addition to arguing that Moss's argument is different in some essential ways from Paul's original one, Pettigrew argues against Moss that it is possible to form justified credences about the likelihoods of having certain preferences following a transformative experience, and that choices for or against transformation can be rational in light of these.

The following two chapters are naturally read as a pair, since both concern whether the novelty of transformative experiences shows that it is impossible for a person to imagine in advance what they will be like. In "What Is It Like to Have a Crappy Imagination?" (Chapter 6), Nomy Arpaly argues that the common failure to understand the subjective perspectives of other people arises from the tendency to simulate these perspectives by imagining oneself in the situation in question—a method that is overwhelmingly likely to fail given the peculiarity of one's own psychology. As Arpaly emphasizes, such failures impede our attempts at interpersonal understanding just as much as they prevent us from anticipating how we ourselves will be different in the wake of personal transformation.

Amy Kind's Chapter 7, "What Imagination Teaches," offers a somewhat more optimistic picture of the powers of human imagination than those of Paul and Arpaly. While accepting there may be some cases where an experience is so qualitatively different from those one has had that it is simply impossible to imagine what such an experience would be like, Kind argues that in the majority of cases a process of imaginative scaffolding can enable a person to grasp the character of an experience that is totally unfamiliar to her. There is, then, no in-principle barrier to using one's imagination to project oneself into a subjectively transformed possible future.

The next three chapters explore in depth several different ways that personal transformation can take place. "Transformative Activities," by Agnes Callard (Chapter 8), reads Elena Ferrante's novel *My Brilliant Friend* as a narrative of personal transformation through a distinctively *active* process of learning. Callard begins by contrasting transformative activity of the sort in question with what she calls transformative *revelation*, in which a person is transformed simply through something that happens to her—whether a chance event or something that she herself has chosen. In transformative activity, not only the onset but the entire process of personal transformation depends on the person's active involvement in learning what she does, and thereby transforming herself into the person she becomes. The protagonists of Ferrante's novel are presented as an illustration of how transformative activity may take place through the *aspiration* to transcend the subjective confines of one's present life.

Nick Riggle's Chapter 9, "Transformative Expression," explores the personally and societally transformative power of participatory art. Riggle's argument focuses on the tension between two common ideas: that an action is *authentic* only if it expresses the agent's dispositions, and that such activity can nevertheless have the effect of transforming the person who engages in it. How can authentic action change who we are when it, by definition, expresses who we are? To address this, he considers several important works from the genre of participatory art which invites participants to express themselves in ways that will alter their core commitments. Riggle draws lessons for theories of action and aesthetic value: sometimes our actions are "authentic" when they issue from a more playful or volitionally open source, and some artworks have aesthetic value in virtue of their inviting such authentic action.

In "Learning from Moral Failure" (Chapter 10), Matthew Cashman and Fiery Cushman consider ways in which the direct experience of moral failure can transform the emotional and motivational contours of one's personality. Cashman and Cushman review a large body of experimental literature that connects moral failure with moral learning, then identify several important conditions in making the experience of moral failure adaptive for the person who undergoes it.

Each of the final five chapters in the volume relates the concept of transformative experience to a further phenomenon which so far has not been considered at length in this connection. In "Risking Belief" (Chapter 11), John Schwenkler asks how we are to think about transformative choices insofar as they promise to convert us—that is, to alter not only our preferences but also our core beliefs and other commitments that are central to our understanding of the world. Schwenkler argues that it can be rational to treat the "doxastically transformative" potential of an experience as a reason to choose against it, but that such a decision must be based in something more than the fact that this experience would alter one's current beliefs. It is when a person knows how things are that she can choose rationally against transformative processes that would destroy this knowledge.

Rosa Terlazzo, in "What Can Adaptive Preferences and Transformative Experiences Do for Each Other?" (Chapter 12), explores the connection between the concept of transformative experience and that of adaptive preferences, or preferences resulting from the fact that a given option was the best available from within a limited set. Terlazzo begins by arguing that employing the concept of transformative experience can help us to see how a person's adaptive preferences can be genuinely beneficial for her, and thus to see such a person as a competent judge of her own good. By means of this argument, she develops an account of adaptive preference with the potential to provide guidance about whether or not to allow ourselves and others to undergo transformative experiences.

Jennifer Lackey's Chapter 13, "Punishment and Transformation," argues that the possibility of personally transformative experience entails that irrevisable, long-term punishments are necessarily irrational. Her argument is that, since the rationality of punishment must be sensitive to the mental state of the person being punished, includ-

ing their mental state after the time of the punishable act, the possibility of radical changes in a person's mental states makes it irrational to punish a person in a way that precludes considering future evidence about those changes. Since strict and long-term punishments, such as life sentences without the possibility of parole, by their nature do just this, such punishments always run afoul of the demands of epistemic rationality.

In "Either/Or: Subjectivity, Objectivity, and Value" (Chapter 14), Katalin Balog argues that the apprehension of value is necessarily subjective, i.e. tied directly to experience rather than to processes of conceptual abstraction. Because of this, she argues, a proper appreciation of value depends on relating to the world through a distinctively contemplative stance that Balog contrasts with the stance of conceptual thought. This contemplative attitude of attentive subjectivity is one that we bring to bear in the appreciation of literature, music, and art. For Balog, it is because of the primacy of subjectivity in the apprehension of value that personally transformative choices cannot be approached through the rubric of rational decision theory.

The final chapter, Evan Thompson's "Death: The Ultimate Transformative Experience" (Chapter 15), considers the transformation involved in the process of approaching one's death. The focus of Thompson's chapter is on the experience of death within the setting of hospice care, which he uses to explore how the existentially transformative experience of dying relates to the experience of one who witnesses a death, in both being with the dying person and bearing witness of her death to others. It is, he suggests, partly because of our own society's tendency to treat death as something private that we have so little insight into the experience it involves.

Acknowledgements

We are especially grateful to Laurie Paul for jump-starting this project, and for guidance and support along the way. The John Templeton Foundation, through funding the Experience Project, supported much of the work in the early stages. Workshops on transformative experience at UNC-Chapel Hill in 2017 and at Yale University in 2018, along with a 2017 pre-conference at the Pacific APA, helped authors share and develop their papers. Peter Momtchiloff has been a wonderful, and patient, editor throughout the process. We are grateful to each author for their work and collaboration.

Enoch's perspective on these issues has been influenced through numerous conversations, reading groups, and talks, and thanks all those who participated and shared their ideas. He is especially grateful for discussions with Laurie Paul, Martin Glazier, John Schwenkler, and Daniel Dennett that helped shape the introduction. He's also thankful for financial support he received while working on this project from the Templeton Foundation, UNC-Chapel Hill, and the Center for Cognitive Studies at Tufts University. He's most grateful to his family.

John's interest in these topics began with a review of Transformative Experience that he wrote for Commonweal magazine. Since then he has learned a lot from conversa-

tions with Kyle Boerstler, Rich Cordero, Enoch Lambert, Marigny Nevitt, Laurie Paul, and Angela Schwenkler, among many others. The Florida State University Council on Research and Creativity supported this research with a grant that paid Erik Franklin to prepare the final manuscript and index.

References

Aumann, R. 1962. "Utility Theory Without the Completeness Axiom." *Econometrica* 30(3): 445—62.

Callard, A. 2018. *Aspiration: The Agency of Becoming.* Oxford: Oxford University Press. Collins, J. 2015. "Neophobia." *Res Philosophica* 92(2): 283-300.

Dougherty, T., S. Horowitz, and P. Silwa. 2015. "Expecting the Unexpected." *Res Philosophica* 92(2): 301-21.

Gigerenzer, G., and R. Selten (eds) 2002. *Bounded Rationality: The Adaptive Toolbox.* Cambridge, MA: MIT Press.

Jackson, F. 1986. "What Mary Didn't Know." *Journal of Philosophy* 83(5): 291-5.

Lewis, D. 1990. "What Experience Teaches." In W. G. Lycan (ed.), *Mind and Cognition: A Reader,* 490-519. Oxford: Oxford University Press.

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

Paul, L. A. 2015a. "Precis of *Transformative Experience." Philosophy and Phenomenological Research* 91(3): 760-65.

Paul, L. A. 2015b. *"Transformative Experience:* Replies to Pettigrew, Barnes, and Campbell." *Philosophy and Phenomenological Research* 91(3): 794-813.

Paul, L. A. 2015c. "What You Can't Expect When You're Expecting." *Res Philosophica* 92(2): 149-70.

Paul, L. A. 2015d. "Transformative Choice: Discussions and Replies." *Res Philosophica* 92(2): 473-545.

Pettigrew, R. 2015. "Transformative Experience and Decision Theory". *Philosophy and Phenomenological Research* 91(3): 766-74.

Simon, H. 1956. "Rational Choice and the Structure of the Environment." *Psychological Review* 63(2): 129-38.

Ullmann-Margalit, E. 2006. "Big Decisions: Opting, Converting, Drifting." *Royal Institute of Philosophy Supplement* 58: 157-72.

Vanderbilt, T. 2016. *You May Also Like: Taste in an Age of Endless Choice.* New York: Alfred A. Knopf.

Velleman, D. 1996. "Self to Self." *Philosophical Review* 105: 39-76.

Von Neumann, J., and O. Morgenstern. 1953. *Theory of Games and Economic Behavior,* 3rd edn. Princeton, NJ: Princeton University Press.

1. Who Will I Become?[(2)]

L. A. Paul

1. Introduction

Life brings opportunities. Opportunities bring change. Sometimes an opportunity is unexpected. You receive a job offer out of the blue. You fall in love. You get pregnant. Other times, it comes after months or years of hoping and planning. You are admitted to the college of your dreams. You decide to get married. You emigrate to a new country.

Such an opportunity can be the chance of a lifetime. It's exciting. You have the chance to discover a new way of living. You have a chance to make something new for yourself, to fashion a new you.

It's also frightening. Change brings risk. Having a baby will change what you care about and how you'll live for the rest of your life. Going to that fancy college will take you into an unfamiliar, challenging new world where you'll have to find new friends and meet high expectations. Moving to a different country means leaving your home and everything familiar behind. Embracing your new love means betraying your partner and destroying your family. Whether you are ready for it or not, having the opportunity to start a new life opens up a whole new world of possibilities, possibilities where you succeed, but also possibilities where you fail.

There are other kinds of life changes we can face. Not all life changes hold promise for the future. Life brings love and friendship and opportunity, but it also brings loss and misfortune. You get divorced. Your sister is diagnosed with late-stage pancreatic cancer. Your son is killed in a car accident.

All of these experiences, good and bad, chosen and unchosen, can be transformative. Transformative experiences are momentous, life-changing experiences that shape who we are and what we care about. By transforming us, they structure the nature and meaning of our lives and the lives of others. They change us, and in the process they reveal ourselves to ourselves, as we recreate ourselves in response to the experience. They make us who we are.

[(2)] L. A. Paul, Who Will I Become? In: Becoming Someone New: Essays on Transformative Experience, Choice, and Change. Edited by: Enoch Lambert and John Schwenkler, Oxford University Press (2020). © L. A. Paul.
DOI: 10.1093/oso/9780198823735.003.0002

Part of the power of a transformative experience is that having it involves discovery. Until you have it, you don't know what it will be like. As you have the transformative experience, something new is revealed to you—what it's like to be in that situation or what it's like to have that experience—and as you discover what it's like, you discover how you react to it. You discover how you'll respond, and in particular, who you become, as the result of the experience.

The way we form and change ourselves through life-changing experiences underlies the way that such transformations define our lives. They shape what we believe and care about, and in this way they make us who we are.

2. Transformative Experience

Transformative experiences, as I define them, affect you in two deeply related ways.

First, they are epistemically transformative: they transform what you know or understand. They do this because they are experiences that are new to you—that is, they are experiences of a new kind, or experiences of a sort that you've never had before, and you have to have this kind of experience yourself in order to know what it's like. By having it, the experience teaches you something you could not have learned without having that kind of experience. When the experience teaches you what that kind of experience is like, and gives you new abilities to imagine, recognize, and imaginatively model possible states involving that kind of experience. Second, such experiences are personally transformative: they transform your preferences. They do this by changing or replacing a core preference, through changing something deep and fundamental about your values. Thus defined, transformative experiences are experiences that change you in both of these ways: they are both epistemically and personally transformative.

Leaving home for college can be a transformative experience. Imagine the moment of departure. Your bags are packed. You've said goodbye to your friends. Your family is waiting at the door. It's time to leave. It's time to start this new part of your life, and you couldn't be more excited. The promise of the open future, of having a world of ideas spread out at your feet, the freedom of having control over your own schedule and your own choices, the thrill of meeting new people and exploring new possibilities: you will stretch your mind in unexpected directions as you enter a new and exciting stage of your life. And a part of you knows that, once you go, you can never come back. Even if you come back, the place will be different. The people will be different. Most importantly, you'll be different. You can return to the place and the people you once knew, but it won't be the same. In this sense, leaving now is leaving forever, because you will never be the same.

Moving to a new place with new challenges, and a new kind of life, confronts you with all the possibility, excitement, and risk that a transformative change can offer. As you prepare to take this momentous, life-changing step, you know that a new life is before you, a life that's very different from the life you've lived up to now. What

will it be like? Who will you meet? What will you do? How will you change? Who will you become? This sense of the open future captures the way the experience will be epistemically transformative: the experience will change you psychologically, but until you are actually there, having the experiences of college life, you can't really know what it will be like. You might know or be told that you are in for these kinds of changes ahead of time, but actually living it teaches you something you couldn't know ahead of time, and in the process, it changes who you are.

Many of life's big personal decisions concern experiences that are transformative in this way. They involve the real possibility of undergoing a dramatically new experience that will change your life in important ways. If you've never done something like it before, the experience will be new and different and mentally expansive, and is likely to change what you care about and how you define yourself. If you are making a decision like this, you face a choice. Should you choose to do it, and discover this new way of living your life? Or should you pass up the chance?

3. Decision-Making

On a natural way of thinking about a life-changing decision, making the choice in the right way is a way of taking charge of your own life. Choices involve responsibility, and to choose responsibly, you need to assess how your choice will affect the world and others in your life. Of course, this is your life we are talking about, and so you also need to assess how your choice will affect you more personally. This is because your choices structure your own life experiences and what happens to you in the world.

An ordinary story of how we are supposed to choose responsibly involves making a rational assessment of the nature of each option. You assess the different possible ways you could act and the different possible results of your act. You map out the different ways the future could develop if you act one way rather than another. You think about what the world could be like, and what you could be like, for each way you could choose. You estimate the value of each path you could take, and the likelihoods of the expected outcomes. Of course, you also take into account expert advice and any moral or social facts that bear on the question of what to do. To choose rationally, you evaluate the options by weighing the evidence and considering the expected value of each act from your own perspective, and then act in a way that maximizes expected value.

How do we go about assessing the values of the options we are supposed to compare when making these sorts of choices? To determine their values and preferences in such cases, people use different types of reasoning, depending on context and previous experience. Often, they rely on one of two basic types of reasoning that are much discussed in the psychological literature: "model-free" or "model-based" reasoning.

Model-free reasoning involves judging options based on retrospective, memorybased assessments (Crockett 2013). When a person reasons this way, they evaluate future

options and possible actions based on cached values they've assigned in the past to similar situations. Such judgments are computationally less demanding. We may be especially likely to reason this way when the possibilities are too difficult or complex to evaluate otherwise.

In contrast to relying on cached values, we can approach decision-making in a different, more expansive way. We can model the hypothetical evolution of our lived experience in response to each possible consequence of each act, in order to assess, judge, and evaluate possible choices. Such "mental modeling" can assist us in developing our point of view in order to determine our preferences concerning the different acts. This is a type of model-based reasoning.

Model-based reasoning is especially useful in high-stakes, deliberative contexts. It is a type of reasoning that

> generates a forward-looking decision tree representing the contingencies between actions and outcomes, and the values of those outcomes. It evaluates actions by searching through the tree and determining which action sequences are likely to produce the best outcomes. Model-based tree search is computationally expensive, however, and can become intractable when decision trees are elaborately branched. (Crockett 2013)

Model-based reasoning is what's most naturally used when you want to deliberate about and carefully assess novel possibilities. You can approach the decision by imagining how you'd respond to different events, reverse-engineering your preferences based on your imagined response. You might start by simulating yourself acting in different ways to bring about different hypothetical options, and then, as you imaginatively represent the outcomes of your actions, assign them value. By imaginatively simulating and assessing your possibilities, you can compare them and determine which act will maximize your expected value.

Your reverse-engineering task could be used to discover your preferences about a possible outcome. However, importantly, it might also be used to create your preferences about a possible outcome. That is, you might need to imaginatively simulate yourself embedded in various events in order to form value judgments about them in the first place, not merely to figure out what you actually prefer given your antecedent values for engaging in those events. Either task involves a form of model-based reasoning.

Using imaginative representation like this can be very useful and important for rational deliberation. We do it all the time when we make ordinary decisions. For example, when you are considering whether you would rather visit a museum or take a stroll in the park, you might start by imagining yourself in the museum, contemplating a series of paintings, in order to assess the desirability of that outcome. You might then imagine yourself walking in the park and admiring the spring flowers, and use your assessment of the appeal of this option to determine your preferences regarding the choice between a visit to the museum and a walk in the park. If you are deciding

whether to go for a swim or to go for a run, you might reflect upon whether it would be unbearably hot to run in the afternoon, while being refreshingly cool to swim. Or it might be numbingly cold to swim in the morning, while being invigorating to run in the cool before the dawn. You plan your daily exercise accordingly.

The same sort of imaginative assessment can be important when we deliberate about major, life-changing decisions for ourselves (or for others). Perhaps you are choosing between two very different types of colleges. If so, you might imaginatively model studying and learning on one kind of college campus, and then imagine doing this on the other kind of campus, and then compare these reflections as you choose where to matriculate. Or perhaps you are pregnant, and you must choose between having the baby or going to college to get an education. To decide, you might imagine life as a parent, with all the joy and sacrifice this entails, and compare it to life as a college-educated person with a wide range of career options. Perhaps, like the French post-impressionist artist Paul Gauguin, you must choose between a life of drudgery and sacrifice where you work to support your wife and family, versus abandoning them for a creative yet self-indulgent life doing what you love. For all of these choices, you might imaginatively model each kind of life you might lead and assess its value, including its moral value, before deciding what to do. For these kinds of life-defining situations, your imaginative assessment of the value of your future life, as well as the future lives of others who will be affected, plays a central role in the way you determine your preferences about which act to perform.

There is a natural connection here with authenticity: roughly, we can think of authenticity in terms of a relation that holds between our current self and our future selves. If I authentically form my future self, my current self intentionally forms my future self as an extension of who I am now, in a way that is consistent with the values of my "true self." Empathetic imagination of one's future self is an important tool that we can use to authentically form ourselves into whom we want to become.

The existence of this sort of reasoning in situations when people face novel, potentially life-changing decisions is documented in McCoy etal. (2019). In this survey, participants were asked about their preferences for a series of high-stakes, fictional options to have transformative experiences (e.g. you are given a one-time only chance to become a vampire, or, a one-time only chance to travel to alien planets). After reporting their preferences, 75 per cent of people sampled from the standard US population and 53% of the philosophers who took the survey decisions indicated that they had learned something about themselves, suggesting that they had discovered something about their preferences by thinking through the novel scenarios.

Imaginatively exploring how we respond in novel scenarios that could be transformative, then, seems to be an important way of discovering how we'd value them, and thus of discovering various truths about ourselves. Such simulation is an important tool to use to discover our preferences when faced with a choice between hypothetical options. Once we assess the values, we can know our preferences and decide how to act.

But there is a problem with relying on your imaginative capacity to envision possibilities for your future self in contexts of transformative choice: that is, when you are choosing whether to undergo an experience that is epistemically and personally transformative. The problem arises when you want to determine your preferences with respect to these novel, hypothetical options.

If you face a decision involving a choice to undergo an experience that will transform you both epistemically and personally, a "transformative decision," you may not be able to imaginatively assess the nature of your future life. This is because you don't have the right sort of epistemic access to your future self. The problem comes from the fact that you can't accurately imagine or simulate what the transformative experience is like. When an experience is a radically new kind of experience for you, a kind you've never had before, you don't know what it will be like before you try it. But you also don't know what you will be missing if you don't try it. You have to actually experience it to know what it will be like for you. As a result, you can't accurately imagine or simulate what it would be like for you to undergo the transformative experience involved. You are in an epistemically impoverished state, facing a distinctive kind of unknown, because you don't know what the experience will be like.

It's a very special kind of situation to be in. In this sort of situation, you have to make a life-changing choice. But because it involves a new experience that is unlike any other experience you've had before, you know very little about your possible future. And so, if you want to make the decision by thinking about what your future would be like if you undergo the experience, you have a problem.

Metaphorically, it's as if you face a blank concrete wall, where you can't see what lies beyond. Perhaps you know that whatever happens in the future, past the wall, will involve you somehow. You know you'll be there, in that future moment, living that future experience. But you don't know what it will be like to be that self. As I will describe it, you face an "epistemic wall."

It's the unknowability that creates the problem, because you can't "see" the outcomes. The basic idea is that if you can't properly represent the points of view of the future selves that are the possible outcomes of your choices, you can't accurately imagine these future lived experiences, and so you can't model them in order to assess how you'd value them as the self who is living that experience. In technical terms, your subjective value function goes undefined for these outputs.

To get a sense of how facing transformation involves facing the unknown, imagine the epistemic situation of a congenitally blind man who is about to gain ordinary vision. Like all of us, his lived experience is formed by his way of experiencing the world through his senses. As a blind person, his dominant sense modality is audition, and thus his way of living in the world is highly defined by his sense of hearing and touch. He has never seen a sunset or watched a movie. This will change when he becomes sighted. Until then, before he gains ordinary vision, there is something he can't know: what it will be like for him to live in the world as a sighted person.

Importantly, descriptions and testimony from others aren't enough to teach him what this is like. Think of admiring the color of the sky just after the sun sets. That color has a particular character, and you couldn't accurately describe what it looks like to him if he'd never had this kind of experience. You could use metaphors, images, and poetry to try to capture its quality by suggesting evocative comparisons, but unless he's already had the right sorts of color experiences, he won't be able to grasp what it is like. For you to be able to accurately describe to him what it's like to experience a sensory quality like light pink, he has to have had the right sort of kinds of experiences beforehand. (And even then, you'd have to describe by using com- parisons––for example, you'd tell him it looks like a shade of pink he's seen before, or maybe like a lighter shade of a color he's already seen.) This is because descriptive language lacks a certain type of expressive power. As a result, some things can only be communicated through experience.

It isn't just sensory experiences that are like this. Many of life's momentous experiences have a special, distinctive character about them, the nature and quality of the experience as lived, that's simply impossible to know about without actually having the experience. It's easiest to see in examples involving the discovery of new sensory qualities, but it isn't confined to them. Other kinds of new experiences can also be like this: for example, living in a world with dramatically new kinds of technology, or experiencing earth-shaking weather events due to climate change, can introduce us to new kinds of lived experiences that we can't know about beforehand.

Even ordinary kinds of experiences, if they are new to you, can be imaginatively inaccessible before you have them. Sometimes this is because they involve kinds or combinations of sensory qualities that are new to you. If the new sensory contribution can't be isolated, or somehow pulled out and separated from the rest of the experience, then to grasp the nature of the lived experience you must actually undergo it, because the sensory contribution is an essential element of the overall lived experience.

For example, think of the distinctive feeling of being in love. Somehow, being in love is made up of a blend of emotion, belief, and desire, and this gives rise to a distinctive kind of experience with a distinctive kind of feeling. Being in love is partly composed of sensation—that is, it consists partly in an experience that has a particular kind of feel or quality, a feel that is inextricably bound up with the experience of being in love. The distinctive nature of the experience of being in love arises, at least in part, from the contribution made by the feeling involved. You couldn't subtract this experiential element out of being in love and still be in love, yet (despite what some popular songs might claim) being in love isn't just a feeling. It's an experience that involves feelings, beliefs, desires, and other rich mental states that constitute the relation you stand in to your beloved. But the phenomenology is still necessary, even if it isn't sufficient. If you've never been in love, you don't know what it is like, and my descriptions here won't be able to teach you. You can know all the things about love that philosophy and science can tell you, and still, when you fall in love for the first time, you'll learn something new. The nature of this complex experience can't be captured with

descriptions any more than descriptions can capture what it's like to see light pink. So if you don't know what it's like to be in love, there's an essential element of the nature and value of being in love that you can't appreciate.

It isn't just the character of experiences like love, fear, awe, and joy that defy description. A person leaving for college can be in the same epistemic boat. They can get descriptions and stories from others about what it will be like for them to start this grand new phase of life, but before they actually leave home and start their new life, there is often a basic and extremely important sense in which they can't know what they are in for.

Moreover, what you discover when you have new kinds of life-transforming experiences isn't just the nature of the new experience. You also discover how you change in response to it, that is, you discover who you become as the result of that new experience. The epistemic transformation changes the way you think and what you care about, and this translates into a new way of understanding yourself and the world around you.

Distinguish between a person persisting over time and the series of selves that realize the person (perhaps realized in turn by a series of more fine-grained temporal parts). On this model, we can think of the new kind of experience as changing a person's life through creating, through the fire of epistemic change, a new self, a new realizer of the persisting individual.

4. Facing the Epistemic Wall

The fact that mere descriptions of experiences can lack expressive power means that when you face a new kind of life-changing experience, you can't rely on descriptions and testimony from others to learn everything you need to know about who you will become as the result of that experience.

You can learn a lot of things beforehand: for example, you can learn a lot of things about falling in love or leaving home to go to college. But actually knowing what the nature of this new kind of experience will be like for you, and by extension what your new lived experience will be like, remains elusive. Because you can't learn from others about what the new kind of experience is like, you can't learn enough to know what it will be like to be the new self that this experience will make you into.

It's the combination of epistemic with personal change that makes this elusiveness worrying. The elusiveness of minor, non-life-changing experiences, like trying a new kind of cereal, isn't something that we worry about. Such experiences are not personally transformative: they don't change who you are. If a new experience isn't a big deal for you, it's easy to skip it or just try it for the sake of discovering what it's like. Trying a new kind of food or reading a new kind of book is like this. If you don't like it, you just move on. If you pass on trying it, it wasn't that important anyway. The epistemic change doesn't scale up into self-change.

A life-changing experience is a much bigger deal. When the new kind of experience is both epistemically and personally transformative, having such an experience is a game-changer. Think about it this way: when the blind man's experience changes in dramatic ways, who he is as a person will also change. But because the experience of becoming sighted will be personally transformative as well as epistemically transformative, his future self is epistemically foreign to him. If he can't know what his future lived experience will be like, there is a deep sense in which he is alienated from his future, sighted self. As I will put it, his epistemic wall generates a selfalienation problem.

The problem of being alienated from one's future and possible selves arises for all transformative choices, both chosen and unchosen. Given the desirability of reasoning rationally when making high-stakes, life-changing decisions, this creates special difficulties for practical deliberation using model-based reasoning. The problem, very simply, is this. You can't know, for some hypothetical future, what it would be like to be the self you'd become in that future. So you can't accurately imagine or simulate this future self. This means you cannot construct an accurate internal model of this lived experience in order to assign it value and determine your preferences. Thus, you've lost one of the main cognitive tools you have for mapping your way through your possible futures and constructing a deliberative response to the choices you face.

Imagine you are facing a choice of whether to undergo a transformative experience. (Alternatively, imagine facing a situation where you are forced to choose between different possible transformative experiences.) In this situation, you don't know what it will be like to have the transformative experience you are making a decision about. This means that you can't accurately imagine or first personally represent what the nature of the lived experience will be like in a way that allows you to imagine who you'll become as the result of the transformation.

If you can't imaginatively represent this possible future lived experience, you can't assess its subjective value—that is, you can't assess the experiential value of the nature and character of this future lived possibility, and thus you cannot determine your preferences. You lack the ability to imaginatively simulate the transformative experience and the future self it could create. As a result, you cannot represent yourself in the way you need to in order to form value judgments about that self or decide which self you prefer to be.

We can draw out the nature of the self-alienation problem in the context of a thought experiment showing how model-based reasoning breaks down when we lack epistemic access to the subjective values for the possible outcomes.

5. Choosing to Have a Child

Imagine yourself in the following situation: you and your partner are trying to decide whether it's time to start a family. In particular, you are trying to decide whether you'd like to have a baby. Your financial situation and physical health make the decision to

become a parent largely up to what you choose—you have the necessary resources, so it's about what you want your future lives to be like. This is a paradigmatic "big decision": the stakes are high, and the choice is irreversible in the sense that, once you've had the child, you can't undo its existence. Even if you give your child up for adoption, you've still become a biological parent.

There are many ways to approach a big decision like this, and, if you have any uncertainty about what you'd prefer, you'll want to think carefully about what you value in order to make the best choice for yourself (and your partner).

Model-based reasoning, where you think about each way you could act, build out the likely consequences of each possible action, and then evaluate and compare these consequences, is the natural way to approach this high-stakes deliberative task. To deliberate, you assess your possibilities to compare them and create (or discover) your preferences. If you can accurately assess the expected value of each act you might perform and compare these values, then when you choose the act that maximizes your expected value, you are choosing rationally.

How are you to assess the expected values of different ways to act? First, you have to be able to assign values to the possible consequences of your actions. Crockett and Paul (forthcoming) show that many people, when they are uncertain about a transformative decision, want to find out what their possible futures would be like. In a study asking people how they would evaluate the possibility of becoming a parent, regardless of whether participants leaned strongly toward wanting children or not, being uncertain about that preference significantly increased the likelihood of wanting to take a (magical) transporter that would allow them to visit a hypothetical future for 24 hours to discover what it would be like to have their child. For those participants who did want children but were uncertain about it, a whopping 96 % of them indicated that they would pay to take the transporter (Figure 1.1).

When a magical transporter to the future is not available, the imaginative simulation of possible futures is a natural substitute.[1] When considering whether or not to become a parent, assessing your possibilities by imaginatively simulating what it would be like for you to have a baby seems like the most natural way to approach the task of assigning value in order to compare alternatives.

The trouble is, for many people, becoming a parent is transformative (Paul 2015b). Having your first child can transform you both epistemically and personally, and for this reason it is an experience that you need to actually undergo in order to know what it is like. When you actually hold your newborn in your arms, you can have an experience like no other, and this can change you in ways that change some of your deepest preferences.

There are, in fact, at least two new kinds of experiences involved: physically gestating and giving birth to your baby, and forming the loving parent–child attachment bond with your child. The second kind of new experience is one that all parents can

[1] See Lewis (1990) for relevant discussion.

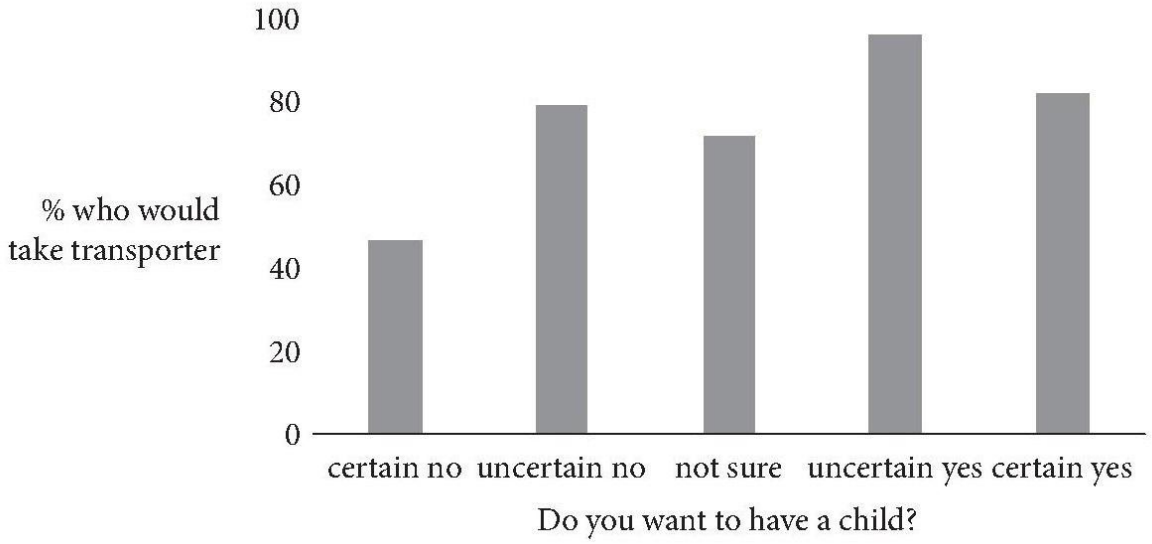


Figure 1.1 Crockett and Paul (forthcoming) found that uncertainty about the preference for having children was significantly correlated with stated preference to take a magical transporter allowing one to visit the future and discover what this is would be like.

have (and for women who have the first kind of experience, it can feed into the second one). These new experiences create a new kind of love in a person, a kind of love that's different from romantic love, and different from the kind of love that you can feel for parents and other family members. It brings intense joy, as well as a new capacity for suffering and vulnerability.

Forming a loving parent–child attachment relation is the source of the foundational shift that parents experience with respect to what matters most to them. Many parents shed their old selves and create new ones, forged by the deep and powerful love they feel for their baby.

The shift involves a core value change. Before they become parents, many people prioritize things like their career, financial success, or social achievements. After they become parents, they care more about their child than anything else. Selfishness turns into selflessness. Put more precisely, most of us have a core preference to pursue our own interests, and this preference is replaced with a preference to pursue our child's interests above our own. This is expressed most keenly by a change in our natural instinct for self-preservation. In a life-threatening situation, while we would want to help others, many of us would save ourselves first, or at least think about saving ourselves first. All of that can change when you have a child. The child (at a level that feels almost instinctual) comes first, even at your own expense. Speaking personally, until becoming a parent, I had never truly understood what it was like to love someone

selflessly enough to be willing, without a moment's hesitation, to sacrifice my life for them. This is just one of the deep and important ways that becoming a parent changed me, and it was a type of change in how I understood myself that I was unable to anticipate until after I became a mother.

The situation is perfectly analogous to that of the blind man before he becomes sighted. The epistemic wall blocks his imaginative capacity to project himself forward into his future self. Given who he is now, he cannot imaginatively see who he will become. In the same way, prospective parents know, before the baby arrives, that a lot is going to change. And yet, they can't know how things will change, in the important experiential sense of truly understanding what their new lives will be like. This means that they cannot prospectively assess their future lived experience as parents. Their epistemic wall blocks their ability to use their imaginative capacities to project themselves forward into their future lives.

If, before you've had your child, you lack the ability to accurately imagine what it will be like, you cannot accurately imagine and assess the nature of this possible lived experience. If you can't do this, you can't use your imaginative abilities to assess the subjective value of this outcome: that is, before you actually have your baby, you cannot assess the value of what it will be like. You can't (accurately) mentally evolve the world forward in order to imaginatively assess what it would be like to have your baby. Therefore, you cannot compare this subjective value to the value of what it would be like to live a child free life. Your subjective value function, which takes as inputs various lived experiences and gives as outputs their respective values, goes undefined at a crucial point.

Epistemic walls create two serious problems.

The first problem, the problem of undefined subjective value, stems from the epistemic transformation involved. It arises for any theory of decision that requires you to assess and compare values in order to maximize your expected utility. If you can't assign the relevant subjective values, you can't compare them to decide which life choice is the best one for you.

The second problem, the problem of self-alienation, stems from the way that the epistemic transformation is the source of the personal transformation. It arises for any view that assumes that rational, reasonable life planning is defined by prospective, informed assessment of one's future possibilities "from the inside," or from the first-person perspective. For many of life's biggest choices, an epistemic wall blocks our psychological access to who we are making ourselves into. A choice to transform becomes, in effect, a leap into the unknown. You make a choice to replace your current self—that is, who you are now—with a new, alien, unknown self.

5.1 Undefined Subjective Value

You might think that the problem of undefined subjective value is easy to solve. Here is one way to solve it for our case of choosing to have a child: forget about

deciding based on what you and your partner's future life will be like. Instead, decide to have a child based on something else, such as whether it will make your mother happy, or because you want to pass on your DNA to future generations. This isn't especially satisfactory, but it is an option.

Another way to solve it is to get friends and relatives with parenting experience to tell you what your subjective values will be (Pettigrew 2015). Now, it's not clear that this is really the right way to get the information that you need. How can someone else know what it will be like for you to have your child? Your mother's experience as a parent is likely to be quite different from what yours would be like. There are further issues here. Your child's nature will have a dramatic effect on your experience as a parent. Before your child is born, can your friends and family members really know enough about what your future child will be like to give you accurate advice about what it will be like for you to have that child? Do we ever really know what our child will be like before we actually produce them? (This may be especially true if your child has a serious mental or physical disability, as the nature of your life as a parent will be significantly affected.)

You could decide instead to rely on what we can know from science and medicine.[2] To determine your expected utility from having a baby, you could draw on the statistical data about happiness and life satisfaction for parents. Such data can't tell you directly what your subjective value for having a child would be. What the data can tell you is what the average effect (or utility value) would be for any member of a population that is composed of individuals like you.

This average effect, however, is perfectly consistent with wide variation in the values assigned to utilities (including the range of uncertainty) for any particular individual member who is included in this average. With real data, we see such variation all the time. At best, you will be able to infer that your subjective value will be within this range.

It's important to see how limited this information is. In particular, you might hope to interpret the average utility values for a member of your population through the filter of your own introspective assessments in order to get a more precise fixon your utilities. This is what people ordinarily do when they want to inform their decisions using scientific evidence. The trouble is that this is precisely what you don't have the ability to do. You can't accurately introspect, because the experience is epistemically transformative. You must simply accept the average utilities, and therefore accept that you are the average person. What this means is that you must replace your own utility assessment with the assessment that applies to the average person. Importantly, you

[2] This sounds appealing, but it's important to note that it's hypothetical at this stage. It's not currently possible for the psychological and social sciences to tell us what your subjective utilities would be for this choice. We have nothing that's even close to good enough data. In fact, for most real-world big decisions at this level, like choosing to have a child, or to emigrate, or to enlist in the Marines, adequately fine-grained statistical data about subjective utilities is unavailable. For a related debate, see my (forthcoming 2020) debate with Paul Bloom in Rivista internazionale di filosofia e psicologia.

are not informing or updating your own prior assessment, because (given that you cannot assess the subjective value) you do not have an informed prior opinion.

Put another way, to choose the act that you expect to have the highest subjective utility by the lights of the average member of your population is not the same thing as to choose the act that you expect to have the highest utility by your own lights. This may be the best we can do in real-world transformative contexts. The value that people attach to the possibility of having a transporter, were such available, to actually visit their future selves suggests that, implicitly, as decision makers, we realize that this is non-ideal.

So the problem of undefined subjective value can be "solved" by eliminating the role of introspection in how you assign your subjective values. You can do this by eliminating subjective value from the decision model, or by assigning yourself a subjective value that is determined solely by the average subjective values of people who are like you in certain ways. "Solving" the problem this way, however, leads us straight into the problem of self-alienation.[3]

5.2 Self-Alienation

Recall that getting testimony about a transformative experience does not teach you what it's like to have the experience: language lacks the expressive power needed to communicate what actually having the experience teaches you. Even if you use the testimony of others to predict the valence and range of your expected subjective value (or if, instead, you dispense with subjective value altogether), when you undergo the transformative experience, you will undergo an epistemic shift that will bring you a dramatically new kind of life. Before you have experienced that epistemic shift, you lack psychological access to your future self. Thus, you lack psychological access to who you are making yourself into until you actually undergo the transformation. As I put it above, you face an epistemic wall.

David Velleman's work on personal identity[4] and persistence brings out the importance of having psychological access to one's future self:

> The future "me" whose existence matters [to me] is picked out precisely by his owning a point of view into which I am attempting to project my representations of the future, just as a past "me" can be picked out by his having owned the point of view from which I have recovered representations of the past. (Velleman 2005: 76)

The way self-alienation arises in transformative contexts can be brought out by exploring two case studies of the transformative choice to have a child.

[3] I don't think these purported "solutions" are at all satisfactory. For discussion see Paul (2015a, 2015b, 2014).

[4] For further discussion see Parfit (1984) and Callard (2018).

In the first case, imagine that you are suddenly confronted with the decision of whether to become a parent. Perhaps your partner unexpectedly gets pregnant. Or (if you are female), perhaps you get pregnant by mistake. As you deliberate about what to do, you ask around for advice.

Pressure to have a child often comes from well-meaning friends and relatives. They say, rightly, that most parents will say they are very happy they decided to do it (Harman 2009). The odds are, then, that you will be happy that you did it. People who know you may also tell you about their own experiences, thinking that your experience will be like theirs. Your friends tell you that they think you should do it because it's the best thing they've ever done. Your mother tells you that she is sure you will be happy if you have a baby.

However, you are not convinced. When you demur or raise worries about the way it could negatively affect your current life, friends and family admit that there are costs, and yet they tell you that once you've become a parent, you'll be willing to make those tradeoffs. You are unimpressed. You tell them that you've seen the haggard parents on the local playground, hair askew, smelling of baby vomit and urine, and that you don't find the thought of being a parent at all appealing. In response, they laugh and explain that you won't mind any of that very much once you actually have your baby.

This is a case where you are to solve the unknown subjective value problem by using testimony from others to determine your subjective values. You are to substitute the judgments of friends and family about your expected subjective value for your own judgment. When their judgments are in alignment with your own, the replacement is easy to make. But when the values they assign are in conflict with your own assessment, the inadequacy of this solution becomes apparent, because it exposes the self-alienation the solution creates.

The self-alienation is obscured when your solution to the value problem lines up with your pre-choice, or "ex ante" beliefs and desires. When there is no value mismatch, and the "ex post" self you become is happy enough, it might not matter much that, after the fact, you couldn't first-personally grasp who you would become. After all, both the ex ante and the ex post selves agree on the choice, and now that it's done, there isn't anyone around to be unhappy about the loss of that old self.[5] But when the self before the choice, the ex ante self, prospectively disagrees with the self that is to be created by the choice (the ex post self) but nevertheless is expected to choose to become that self, the alienation of the ex ante self from the ex post self becomes apparent (Paul and Quiggin 2018).

The issue is not that the judgments of your friends and family members are incorrect. Assume their judgments about your future subjective values are in fact correct. (If their judgments are false, you haven't solved the unknown subjective value problem.) The trouble is that, at the time of choosing, their judgments about the best choice

[5] And as luck would have it, the human psyche is enormously successful at adapting its preferences to be happy with whatever outcome it finds itself in.

to make conflict with your judgment. The choice that is deemed "rational" by their testimony cuts deeply against what you want and believe now. This means that, to choose rationally according to your friends and family, you should choose to become a parent, even though you don't want to. The lesson is that, to be rational, it doesn't matter what *you* think. What matters is what the people who know you (especially the people who are already parents) think.

This seems bad. But why? After all, choosing rationally is the choice that maximizes your expected value. Their testimony guides you to the rational choice. Why isn't this obviously and uncontroversially the best thing to do?

Part of what's bad is that you are trading in your autonomy for the sake of your rationality. Your solution to the unknown subjective value problem—one that relies on others to tell you the subjective value of your future life—eliminates an important part of your role in your value assessment. For one of the most important and personal decisions of your life, you are forced to rely solely on the judgments of others.

The real problem, however, isn't merely that you have to rely on others to make one of your most important personal life choices. That's just what leads to the real problem. The real problem is that, to be rational, you must make this choice by rejecting what you care about. In order to choose rationally, you must to choose to become a self that is alien to who you are now. Solving the value problem through relying on testimony from others creates the self-alienation problem.

The self-alienation you face, at its root, stems from the deep epistemic change you will undergo. When there is a value mismatch between your ex ante and your ex post selves, the solution to the unknown subjective value problem alienates you from your choice and, by extension, from the self you are choosing to become. The self you will become is first-personally foreign to who you are now. And it is because this self is so alien that you face the unknown subjective value problem in the first place: you can't imagine yourself into that self's perspective in order to assign value to that new lived experience. The deep epistemic change of the transformation is the common source of the unknown subjective value problem and the self-alienation problem, and this is why merely solving the value problem won't eliminate the alienation problem.

We can explore a second case in order to draw out the formal structure of the selfalienation problem. In the second case, we flip the results. Instead of being skeptical, you find yourself incredibly keen to become a parent. You've read novels and seen films about how joyful and satisfying it can be to have a family. Your sister just had a baby and she can't stop talking about how happy she is. You've always wanted to become a parent, and feel that having a baby would make your life fulfilling and meaningful.

However, your friends and family counsel you to avoid becoming a parent. Perhaps they think that you and your partner are not ready, or that you would be unhappy as a parent. When you consult the scientific evidence, it also goes against you: it tells you that the average subjective value for people like you is negative. The evidence suggests that the quality of your future life will decline if you become a parent. Again, to choose rationally, you must substitute the judgments of others in place of your own.

Let's assume that you believe in the science, and in the wisdom of your friends and family members. So you accept their assessment, even if it does not comport with what you believe about how you would respond. As a result, you are epistemically alienated from your rational choice by your imaginative incapacities. This is simply the bargain you must make in order to solve the unknown subjective value problem.

Let's add detail to see how the reasoning might go. You have precise credences for each of the relevant hypotheses and their associated outcomes: having a baby versus not having a baby. You consult the best scientific sources available, and the research clearly tells you that you can expect a low utility if you have a child, and a high utility if you don't. (Friends and family agree with this result.) In effect, the science tells you that you will maximize your expected utility by choosing to remain childless, even if you are uncertain about just how much.

You can't understand this intuitively, because although you don't feel like you have a detailed grasp on what the future would be like (everyone tells you to expect a dramatic life change), your own assessment of your utilities for having a child by imaginatively or introspectively prefiguring your future self as a parent assigns a very high utility to having a child, and a very low utility to not having one. In short, you desperately want to have a child, and it "feels right" to you to have one.

Given that it's not rational to choose to act in a way that does not maximize your utility, then according to the expert's assessment of your utilities, you can't rationally choose to have your child, even though this conflicts with your personal assessment. In this situation, to be rational, you must allow the expert determination of what you are to believe about your utilities to replace your introspective assessment of your heart's desires.

How does the evidence predict your expected utility? Let's walk through the way the counterfactuals work. To assess your utilities in different possible circumstances, we start by considering you in the actual world, @, at ti, and then assessing your utilities at t2 in different possible worlds Wi and W2. In Wi at t2, you have a baby, and in W2 at t2, you do not have a baby. We assess your utilities in different possible worlds because we are assessing what the actual world would be like under different possible changes of state, viz. having a baby or not having a baby. (Recall that, before you have a baby, because of the transformative nature of the experience, world Wi at time t2, where you have a baby, is epistemically inaccessible to you. However, we assume the science tells you what utility to expect in those circumstances.)

Do you exist in Wi and W2? Yes—or at least your respective counterparts do. Call the person who exists in Wi at t2, "Ci" and the person who exists in W2 at t2, "C2." The self-alienation problem rises with Ci, the person who is identical to you (or the counterpart who represents you) in Wi at t2.

Here is the root of the problem. Normally, when we counterfactually assess the value of a state change for an agent A, the salient dispositions and preferences of A are kept fixed in order to assess A's proposed utility in the new state. In other words,

we preserve *act-state independence.*[6] After all, at ti, we are considering what A wants, and trying to assess whether a potential change in circumstances (a change in the state of the world) suits A's preferences. If we are interested in what the value of the change would be for agent A, then we want to compare A (ex ante) in their current circumstances to A in their new (ex post) circumstances. If A's preferences also change when there is a change in circumstances, then at t2 we aren't getting information about whether the proposed change suits A's current (ex ante) preferences.

In other words: when you prospectively assess what the value of a change in circumstances would be for you, you want to compare yourself in your current circumstances to your counterfactual self in your new (changed) circumstances. But if *who you are also changes* in the new (counterfactual) circumstances, finding out your future utilities (from science, or via testimony) in those future circumstances doesn't tell you whether the change fits who you are right now ex ante (that is, at the time of choosing). The problem is that act-state independence has been violated.[7] Example: I might be happier if I had a frontal lobotomy tomorrow. After the lobotomy my ex post self might even testify to my new preference to have been lobotomized, finding it quite pleasant indeed. But right now, I (ex ante) definitely don't want to get a lobotomy! I feel quite sure I, ex ante, should disregard the preferences of that hypothetical lobotomized ex post self. That prospective future self's testimony is not relevant to me, precisely because act-state independence has been violated in the creation of that self.

Choosing to have a child puts you in an interestingly similar position.[8]

For the change modeled by Wi, becoming a parent, act-state independence *fails.* This is because, by hypothesis, the state change represented by Wi at t2 (you with your baby) does not exist in isolation. The change you undergo by having a child is transformative. That is, changing the state of the world to make you into a parent would *also* change your preferences and your psychological capacities. If you are (or are represented by) Ci in Wi at t2, in Wi you are a person with a radically changed first-person perspective.[9]

We can put it this way: at ti, in @, when you consider the choice to have a baby, from your first-person perspective, Ci's point of view is psychologically alien to you.[10] You cannot project your point of view into Ci's point of view, or grasp her point of view as an extension of your own.[11]

---

[6] This is often characterized as an "independence" axiom or theorem.

[7] For a developed discussion of this problem, see Paul and Healy (20i6).

[8] See Barnes (20i5).

[9] It represents a change the features of the agent whose utility is being assessed, not just the circumstances of the world in which the agent is embedded.

[10] Or, we might say, Ci isn't who you, from your @-at-ti vantage point, would identify as your psychological counterpart.

[11] This brings out an interesting mismatch between how we think about personal identity from an impersonal, bird's-eye point of view, and how we think of it from within. Bernard Williams (i97o) captures this mismatch in his discussion of conflicting intuitions in his "The Self and the Future."

While Ci might be, strictly speaking, personally identical to you, from your actual perspective at ti, Ci is not an eligible future self, because Ci is not psychologically accessible to you in any first-personal sense.[12]

So the utilities that the science and the anecdotal testimony predict you'll have in Wi are not the utilities of anyone you can recognize as your future self. They are indeed the utilities of Ci at t2, but from your first person perspective at ti, Ci is *not you.*[13] When you consider your decision at ti, you want to know how you'll respond to the experience at t2–that is, whether your preferences will be satisfied. Wanting to have *your* preferences satisfied carries with it an implicit, psychological, first-personal constraint: when you make an important personal decision, you want to know the (range of) utilities that the person who you can first-personally identify as your future self will have.

In other words, when you assess your choices, you want to have psychological access, in an anticipatory or imaginative way, to each of your possible future selves. For each possible choice, you want to grasp the first-person perspective of the self who you think you could become, and who will live with the result of your choice.[14]

Because, from your first-personal perspective at ti, Ci is not you (or, if counterpart theory is preferred, we can say that from this perspective Ci is the wrong counterpart), relying solely on testimony to tell you about your future utilities creates alienation from your possible future selves. You are psychologically alienated from who results from this change.

(Perhaps in the strict metaphysical sense you are the same person after having your child, just like you are the same person now as you were when you were three years old. But in another, very important sense, you aren't the same person—that is, you are not the same self, and this may be what people mean when they say "I'm not the same person I used to be." The metaphysics here involves distinguishing "same self" from "same person." Think of a person as composed of a series of selves over time, and of those selves in turn as composed, most fundamentally, of a series of temporal stages, or temporal parts. We can then mark different requirements for being the same person in some literal or strict sense at different times and for being a different self over time (and further, we can distinguish between different temporal parts of the same self, or of the same person, at different times). The important distinction, then is between being very different selves over time and being literally a different person at different

[12] On some metaphysical accounts of personal identity, the *metaphysically same person* relation merely requires the right sorts of causal or other sorts of continuity. The point here is that *metaphysically same person* and *same self* are different relations, and the one that matters in these decision contexts is the same self-relation. We can think of this as a problem of personal ambiguity, as opposed to a problem of personal identity.

[13] Or, I'd be inclined to say, Ci is the *wrong* counterpart. "It's the wrong trousers, Gromit, and they've gone wrong!"

[14] Counterpart theoretically: you want to know the (range of) utilities of a counterpart that is psychologically similar in the relevant first-personal sense to who you are now.

times. Strictly speaking, I'm still the same person: after all, I have the same birth certificate and the same parents. I'm just not the same self. I've changed in core ways, and colloquially, that's usually what we mean by saying "I'm a different person." In this way, a person can be literally the same person over time while also being composed of a chain of changing selves, which is in turn composed of a chain of temporal parts. Each link of the chain is a different part, and the links summed together compose the person over time.[15])

The point of all this isn't that you shouldn't have a baby. Maybe you should. The point is that choosing to have a child, like choosing to become sighted after a life of blindness, involves facing an epistemic wall, and the implications of this are personally significant. Having a baby can be a transformative experience, and so choosing to become a parent can mean you are choosing to become a different kind of person. You are choosing to become a self who is unknown to you now.

And this is an extraordinarily salient implication of transformative experience and transformative choice. You might think that, when you undergo a transformative change like becoming a parent, you are just realizing a future version of who you are now. But you aren't: you are *replacing* who you are now with some radically different, alien self. That self is you, in some sense. But not in any first-personally accessible sense.

Drawing out the nature of the choice in this way can help us to understand some of the implications of the first case we considered, where you don't want to become a parent, but everyone advises you to have a baby. In that example, testimony from those who have children seems to suggest that, even if you can't really understand what it's like to be a parent, you should do it anyway, because you'll maximize your expected utility. In that case, you admit that parents are in general happy with their choice, and you even admit that, if you became a parent, you'd probably be happy with your choice, even if, right now, you don't want to be a parent.

In such a case, what should you infer? That is, what if your friends are right that, like them, you'd be very happy and satisfied as a parent? What if your mother is right that, after having your baby, you won't care about all the things you care about now? Does this mean that, deep down, you'd really prefer to be a parent, even if you can't know it now? Does this mean that your mother really does know you better than you know yourself?

The suggestion seems to be that your friends and family members understand something you don't. Once you become a parent you'll understand that they were right all along. In more technical terms, the implication seems to be that, once you have your child, your underlying preference to become a parent will be revealed. You can't understand this ahead of time, but that's because the experience is epistemically transformative.

[15] For relevant discussion see Parfit (1984), Paul (2017), and Pettigrew (2020).

I reject this. To start: why is this sort of well-meaning advice so irritating? It's even more upsetting when science is brought to bear. When the scientific evidence supports a choice that you intuitively reject, the implicit suggestion seems to be that resisting it involves some sort of magical thinking. The suggestion is that you are confused: you think you are unique or that the science doesn't apply to you, but if you had a proper understanding of the way empirical evidence worked, you'd know better. Aspersions are thus cast on your ability to understand the nature of statistical inference or to make rational decisions.

What is it that feels so objectionable here? Is it merely the smug paternalism of such suggestions? Are you simply wrong, and unable to face this fact? No.

Again, assume your friends and family members are, in fact, correct in the conclusion that, if you become a parent, you'll be happy with your choice. And assume there is scientific evidence to support this conclusion. But this fact, considered alone, does not mean that you should choose to become a parent. The philosopher Elizabeth Harman (2009) has argued persuasively that "I'll be glad I did it" reasoning can fail in some cases. This seems to be such a case. (Harman might not agree—see her 2015.)

We can see the mistake when we focus on the transformation involved. Recall that, under the state change we are considering, act–state independence is violated. Having a child is not just epistemically transformative: it is also personally transformative. Becoming a parent transforms what matters to you, and can make you happy and satisfied to be a parent. This transformation can happen even if, before you have a child, you really, truly, don't want to become a parent. So, again, does this imply that when you have a child, your underlying preference to become a parent is suddenly revealed?

No. There is another, very plausible explanation that fits your situation: something about becoming a parent eliminates your old preferences and implants new preferences in you. When you are transformed, your preferences are transformed. What you care about is transformed by the process itself. Even if you don't want to become a parent, the process of forming an identity-defining attachment to your child can create or implant new preferences, replacing your old pre-kid cares with new kid- focused ones. What a person cares about can change, hugely, when they have a child, and this happens in virtue of the psychological and biological changes that make them a parent.

If so, then your concerns about the choice are perfectly legitimate. You are not being perverse. You are not confused. You are not ignorant of your own preferences. Your worry is not about whether you'll be happy with who you've become *after* you've been transformed.

Your worry is that, right now, what you care about—now—isn't consistent with being transformed. Becoming a parent would change you in ways that, right now, you reject. If you do not want to have a child, then, in your current childless state, you don't care about the things you'd care about as a parent, and, even more importantly, you don't want to care about them. You want to preserve who you are *now,* and what you care about *now.* In these circumstances, it's perfectly reasonable to resist the pressure

you are getting from the experts. That's because there is no implication that somehow, becoming a parent would be better for the self you are now. Rather, becoming a parent would *replace* the self you are now with a different self, an alien self: a self that, right now, you don't want to become.

So the well-meaning advice from friends and family is too simplistic. Their advice is flawed, because it does not account for the true structure of the problem. For the same reason, the scientific evidence fails to apply in the clean way that the results might suggest. The clean application assumes that act-state independence is preserved. But when act-state independence is violated, it is unclear how one should interpret the statistical results, and thus it is unclear what one can infer about what is rational to choose in these circumstances.[16]

What we have here is a first-personal version of a Kuhnian revolution. In transformation, we replace our old point of view, our self-understanding of who we are, with a new, incommensurable point of view, a new self-understanding of who we are. Instead of a conceptual revolution writ large, like that brought on by the discovery that the earth revolves around the sun (which replaced the old idea that the sun and other planets revolve around the earth), we experience a conceptual revolution writ small.

The point can be made another way. When you face a transformative experience, even if friends, family, or science can tell you about the utilities involved, you still face an existential problem, one that they are unqualified to address: Will *you* be happier after the transformative change? Or will you just become someone else?

References

Barnes, E. 2015. "What You Can Expect When You Don't Want to Be Expecting." *Philosophy and Phenomenological Research* 91(3): 775-86.

Callard, A. 2018. *Aspiration: The Agency of Becoming.* New York: Oxford University Press. Crockett, M. J. 2013. "Models of Morality." *Trends in Cognitive Sciences* 17(8): 363-6.

Crockett, M. J., and L. A. Paul. forthcoming.

Harman, E. 2009. ' "I'll Be Glad I Did It": reasoning and the significance of future desires'. In J. Hawthorne (ed.), *Ethics,* 177-99. Hoboken, NJ: Wiley-Blackwell.

Harman, E. 2015. "Transformative Experiences and Reliance on Moral Testimony." Res Philosophica 92(2): 323-339.

Lewis, D. 1990. "What Experience Teaches." In W. G. Lycan (ed.), *Mind and Cognition,* 29-57.

Hoboken, NJ: Wiley-Blackwell.

[16] Paul and Healy (2016).

McCoy, J., T. Ullmann, & L. A. Paul. 2019. "Modal Prospection." In A. Goldman & B. McLaughlin (eds), Metaphysics and Cognitive Science. Oxford University Press, 235-267.

Parfit, D. 1984. *Reasons and Persons.* Oxford: Oxford University Press.

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

Paul, L. A. 2015a. "Transformative Choices: Discussion and Replies." *Res Philosophica* 92(2): 473-545.

Paul, L. A. 2015b. "What You Can't Expect When You're Expecting." *Res Philosophica* 92(2): 149-70.

Paul, L. A. 2017. "The Subjectively Enduring Self." In I. Phillips (ed.), *The Routledge Handbook of the Philosophy of Temporal Experience*, ch. 20. Abingdon: Routledge.

Paul, L. A., and K. Healy. 2016. "Transformative Treatments." *Nous* 52(2): 320-35.

Paul, L. A., and J. Quiggin. 2018. "Real World Problems." *Episteme* 15(3): 363-82.

Pettigrew, R. 2015. "Transformative Experience and Decision Theory." *Philosophy and Phenomenological Research* 91(3): 766-74.

Velleman, J. D. 2005. *Self to Self: Selected Essays.* Cambridge: Cambridge University Press.

Williams, B. 1970. "The Self and the Future." *Philosophical Review* 79(2): 161-80.

2. Being Someone Else[(3)]

Martin Glazier[(4)]

1. Introduction

I have sometimes wondered why I am who I am. Why, out of all the people in the world, am I this one? Is there not something arbitrary in the fact that I turned out to be Martin Glazier rather than someone else? Couldn't things have been otherwise?

Many have been tempted to answer "yes." Thus David Lewis (1986) writes:

> Here am I, there goes poor Fred; there but for the grace of God go I; how lucky I am to be me, not him. Where there is luck there must be contingency. I am contemplating the possibility of my being poor Fred, and rejoicing that it is unrealized. (p. 231)

Lewis was not alone. Thomas Nagel (1986) writes:

> My being TN (or whoever in fact I am) seems accidental So far as what I am essentially is concerned, it seems as if I just happen to be the publicly identifiable person TN—as if what I really am, this conscious subject, might just as well view the world from the perspective of a different person From a purely objective point of view my connection with TN seems arbitrary. (pp. 60-61)[1]

And in a similar vein, Bernard Williams (1973) writes:

[1] This passage and the one below are quoted in Ninan (2009: 447).

[(3)] Thanks to Chris Blake-Turner, Nilanjan Das, Kit Fine, Rachel Fraser, Thomas Hofweber, Enoch Lambert, Kathryn Lindeman, Annina Loets, Kris McDaniel, Carla Merino-Rajme, Daniel Nolan, L. A. Paul, Zee Perry, John Schwenkler, Ted Sider, Olla Solomyak, Bart Streumer, Jennifer Wang, and Susan Wolf, as well as to audiences at Cologne, Leeds, North Carolina, and the Pacific APA. I am grateful for the support of the John Templeton Foundation.

[(4)] Martin Glazier, Being Someone Else In: Becoming Someone New: Essays on Transformative Experience, Choice, and Change. Edited by: Enoch Lambert and John Schwenkler, Oxford University Press (2020). © Martin Glazier.

DOI: 10.1093/oso/9780198823735.003.0003

> "I might have been somebody else" is a very primitive and very real thought; and it tends to carry with it an idea that one knows what it would be like for this "I" to look out on a different world, from a different body, and still be the same "I". (p. 40)

These philosophers have been attracted to the "contingentist" thought that it is possible that I should be someone else.

Contingentism is also presupposed by those philosophers who take it to be possible that I should become someone else. L. A. Paul, for instance, has argued that some of life's biggest choices, such as the choice to become a parent or to pursue a certain career, can be "transformative." A choice that is transformative, in Paul's sense,

> involves the possibility of undergoing an experience that changes you from the self you are now into a different, new self When confronted with a transformative choice, you must decide whether to replace your current self and its perspective with a new self and that self 's perspective. (Paul 2015: 490)

Although Paul here glosses transformative choice in terms of the notion of the self, it is not altogether clear from this passage how she wishes to understand that notion. But what is clear is that she thinks there can be distinct things s and s^l such that although I am s, I will later be s^l . In this sense, she thinks it is possible that I should become someone else. But it can hardly be the case that I will be s^l unless it is possible that I should be s^l . And so Paul must take it to be possible that I should be someone else. Indeed, anyone who understands transformative choice to involve becoming someone else is in this way committed to contingentism.[2]

Despite these philosophers' attraction to contingentism, however, the view may seem absurd. For how could I be someone other than who I am?

The aim of this chapter is to show that contingentism is, if not true, then at least not absurd. I will argue that the tenability of contingentism may be upheld—in fact, can only be upheld—by appeal to a notion of perspective that is distinctively metaphysical, in contrast to the more usual epistemic notion. It is in terms of this notion of perspective that the contingentist may distinguish two senses of metaphysical possibility, one in which I must be MG and one in which I might be someone else.

Although I have spoken about myself and have asked about the possibility that I should be someone else, in Cartesian style my considerations are intended to apply to you as well. As you read this chapter, you should translate the arguments I give about myself into arguments about yourself.

[2] Of course, there may be other ways of understanding transformative choice that are not so

2. The Challenge to Contingentism

I begin by clarifying the contingentist view and the challenge it faces. The contingentist claims that it is possible that I should be someone else—poor Fred, say. But how could I be someone other than who I am?

The contingentist's claim must not be misunderstood. Although she does think it possible that I should be someone other than who I am, she of course does not think it possible that I should both be Fred and also not be Fred. Her position is rather that although I am not Fred, it is nonetheless possible that I should be Fred.

There is a more subtle way in which the contingentist's claim might be misunderstood. It should not be taken to be a de re claim of possible identity—a claim of the form "a = b." We should not understand the contingentist, that is, as saying of Fred and me that it is possible that we should be identical. For Fred and I are distinct objects, and objects that are distinct are necessarily so.

Let us call a statement of the form "a = b" an identification. The contingentist must recognize a non-identificational reading of statements of the form "I am Fred."[3] It is on this reading that she takes it to be possible that I am Fred. Indeed, I believe it is such a possibility that Nagel has in mind in considering whether he "might just as well view the world from the perspective of a different person" and that Williams has in mind in discussing "what it would be like for this 'I' to look out on a different world, from a different body." Caspar Hare, a recent proponent of contingentism, also appears to understand the view in this way. He glosses the possibility that I should be Ralph Nader as the possibility that I should "see the world through Ralph Nader's eyes" (Hare 2009: 82). I will have to leave for another time the question of how precisely this non-identificational reading is to be understood. But these philosophers' remarks give enough indication of its meaning for our purposes.

This distinction between the identificational and non-identificational readings of "I am Fred" might be thought to provide an adequate response to the challenge to contingentism. For the contingentist may agree that objects that are distinct are necessarily so and yet insist that it is nonetheless possible that I am Fred when "Iam Fred" is given its non-identificational reading. Of course, no reasonable response to the challenge will judge it to be wholly misplaced. There is surely something absurd in the vicinity of contingentism, even if it is not the view itself. But the contingentist can acknowledge this. For she may concede that it is not possible that I am Fred when "I am Fred" is given its identificational reading.

In fact, however, this response is not adequate. For even on the non- identificational reading there is a point of view from which it does not seem possible that I am Fred. To see this, notice first that the statement "I am MG" is true on this reading. After all, if we are willing to state the possibility that I am Fred by saying that I might just

committed.

[3] Cf. Nagel (1986: 55): "From this point of view it can appear that 'I am TN,' insofar as it is true, is notan identity but a subject-predicate proposition."

as well view the world from the perspective of Fred, or through Fred's eyes, then since I do view the world from the perspective of MG, and through MG's eyes, we should admit that I am MG.

What is more, not only am I now MG, I have been MG my whole life. And this is no accident. It is no accident that every morning I wake up and once again find that I view the world from MG's perspective. How could it be otherwise? That is just who I am! From this point of view, then, it seems necessary that, on the non- identificational reading, I am MG, and therefore not possible that, on this reading, I am Fred. The challenge to contingentism remains.

Since the contingentist must state her position using only the non-identificational reading, and since the challenge to contingentism is in turn stated using only that reading, let us restrict ourselves from now on to the non-identificational reading of statements of the form "I am Fred" unless we explicitly say otherwise. The contingentist's position is then that there is someone else s—someone s such that it is not the case that I am s—such that it is possible that I am s. The challenge is that there is a point of view from which this does not seem possible.

It might be suggested that the contingentist should retreat from the letter of her position. Perhaps she should simply reject the possibility that I am someone else, even on the non-identificational reading. What is really possible, she might say, is not that I am Fred but that I am in just the situation or predicament Fred is (actually) in. This "surrogate" possibility claim, it might be argued, avoids the challenge while preserving the spirit of the contingentist view.

If this surrogate response is to have any hope of success, the surrogate possibility claim must be properly understood. It cannot be taken to be the claim that it is possible that what is true of me should be just what is actually true of Fred, or, to put it another way, that it is possible that I should satisfy just those open sentences that are actually satisfied by Fred. For it is actually true of Fred that he is Fred and that he is not MG. But according to the surrogate response, it is not possible that I should be Fred or that I should fail to be MG.

The surrogate possibility claim must therefore be understood in some weaker fashion. A natural thought is to understand it as the claim that it is possible that I should satisfy just those qualitative open sentences that are actually satisfied by Fred. But I do not think this version of the surrogate response preserves the spirit of the contingentist view. Although in the case of Fred this fact may be hard to see, in other cases it is manifest. Suppose for instance that I inhabit a universe with twofold (180°) rotational symmetry. Here I stand, one meter from the axis of symmetry, staring down my doppelganger.[4] The contingentist will think: I could have been him! According to the surrogate response, what is really possible is not that I should be my doppelgänger, but that I should satisfy all and only the qualitative open sentences he satisfies. But

[4] Alternatively, we might suppose I inhabit a world of two-way eternal recurrence, in which history repeats itself endlessly with no first epoch and no last epoch.

since the universe is symmetric, we may plausibly suppose that I already satisfy these sentences. Yet it is part of the contingentist's thought that something further is possible, something which is not actual: that I am my doppelgänger. The surrogate response does not do justice to the contingentist's sense of unactualized possibility.

In this case, then, the contingentist admits that not only is it possible that I should be in just the predicament my doppelgänger is in, it is also possible that I should be him. She should do the same in the case of Fred. She should insist that not only is it possible that I should be in just the predicament Fred is in, it is possible that I should be Fred. She should therefore reject this version of the surrogate response.

Have we considered too weak an understanding of the surrogate possibility claim? It might be thought that the lesson of the symmetric universe case is simply that the merely qualitative is not enough: the contingentist must require that I satisfy more than just the qualitative open sentences Fred satisfies. Of course, we have seen that she cannot require that I satisfy all the open sentences Fred satisfies. But might there be some intermediate requirement that will serve?

It has been suggested to me that the contingentist should understand the surrogate possibility claim as the claim that it is possible that I should satisfy just those open sentences that Fred (actually) satisfies, provided that they do not involve either of us. But this version of the surrogate response does not seem to me to be adequate either. To see why, suppose there is nothing in the symmetric universe other than my doppelgänger and me. (Not even points of space, if we are suitably relationalist.) As I confront my doppelgänger in that inky void, the contingentist will think: I could have been him! But he and I are the only occupants of the universe and so we may plausibly suppose I already satisfy all and only the open sentences not involving either of us that he satisfies. Yet the contingentist will take something further to be possible: that I am my doppelgänger. I therefore think that this version of the surrogate response also fails to preserve the spirit of the contingentist view.

Is there some more subtle version of the response that succeeds? It is hard to be definitive here, but a case can be made that any version will face the following dilemma. Either the proposed understanding of the surrogate possibility claim is such that if I satisfy the condition specified in the claim then I am not MG, or it is not. If the former, it is inadequate: the surrogate response, after all, is supposed to avoid the conclusion that it is possible that I should fail to be MG. And if the latter, then I believe it will be possible to construct a case in which my doppelgänger and I both satisfy the condition. Upon considering this case, I believe, it will be clear that the contingentist will think that something further is possible: that I am my doppelgänger. And so the response will be seen to be inadequate.

A different response to the challenge involves distinguishing senses of "I." It will be suggested that the contingentist should take it to be possible in one sense of "I" that I should be Fred, but take this to be impossible in another sense of "I." What might these senses of "I" be? Philosophers have offered a number of suggestions. For instance, Descartes distinguished the body from the soul. Kant distinguished "phenomenal" and

"noumenal" selves. And the early Wittgenstein distinguished the human being from the "metaphysical subject."

Further suggestions continue to emerge in the contemporary period. For example, one might distinguish the "center" of a Johnstonian "arena of presence and action" from the human being who occupies that center.[5] One might distinguish the Nagelian "objective self" from the "particular person" through whom this objective self views the world.[6] Or one might distinguish between Finean "metaphysical" and "empirical" selves.[7]

Although there is some difficulty in understanding these potential senses of "I," they are not without interest. But they do not deliver a successful response to the challenge to contingentism, since it will arise again even if these senses can be distinguished. To see this, take whatever sense of "I" is supposed to vindicate the possibility that I should be Fred; to fix ideas, let it be my Cartesian soul s. Of course, mine is not the only soul in the world; there are others as well. The contingentist will surely wonder: why, out of all the souls in the world, am I *this* one?[8] And if s^l is a soul other than mine, the contingentist will surely find compelling the following Lewisian thought:

> Here am I, there goes poor s^l; there but for the grace of God go I ... I am contemplating the possibility of my being poor s^l, and rejoicing that it is unrealized.

The contingentist, then, will take it to be possible that I should be s^l . Yet I may at the same time reflect, "I am *s* and have been my whole life. Indeed, it is no accident that every morning I wake up and once again find that I am s. How could it be otherwise?" There is still a point of view, then, from which it seems necessary that I should be s and thus not possible that I should be s^l . The problem has not been resolved, only relocated.[9]

The contingentist might think to deny that it is possible that I should be s^l , thus leaving her free to concede the necessity of my being s. But she takes it to be possible that I should be Fred. How can she accept this possibility while denying the possibility that I should be s^l ?

She might try to drive a wedge between the two in the following way. We have supposed that the contingentist's response to the original challenge involves taking it to be possible that I should be Fred when "I" refers to my soul s. And so she might insist that it is not possible that s should be s^l and thus not possible that, in this sense of "I," I should be s^l.

---

[5] Johnston (2010: ch. 2).

[6] Nagel (1986: ch. 4).

[7] Fine (2005).

[8] Here and elsewhere in this chapter, we take the contingentist to be capable of thinking about me in the first person. Perhaps the figure of the contingentist is therefore best understood as "me with my contingentist hat on" or "me in my contingentist moods."

[9] Cf. Lewis (1986: 232).

But whether or not this is correct, it does nothing to show that there is not some sense of "I" in which it is possible that I should be s^l. And the contingentist faces considerable pressure to say that there is such a sense. After all, she finds it compelling that I might contemplate the possibility of my being Fred and rejoice that it is unrealized. Why is it any less compelling that I might contemplate the possibility of my being s^l and rejoice that it is unrealized? Or again, the contingentist finds it compelling that I might just as well view the world from a perspective other than that of MG. Why is it any less compelling that I might view the world from a perspective other than that of s?

But if the contingentist admits that, in some sense of "I," it is possible that I should be s^l, then she will face the challenge. For, as before, there will be a point of view from which it seems that, in this same sense of "I," it is necessary that I should be *s* and thus not possible that I should be s^l .

A related response to the challenge involves distinguishing senses, not of "I," but of "Fred." One might think the contingentist should distinguish Fred's body, Fred's soul, Fred's phenomenal self, Fred's noumenal self, and so on. She might then take it to be possible, in one sense of "Fred," that I should be Fred, while conceding that in another sense of "Fred" this is not possible. But here the challenge will arise once more in much the same way as before. To see this, take whatever sense of "Fred" is supposed to vindicate it's *not* being possible that I should be Fred; to fix ideas, let it be Fred's soul *s*'. Just as before, it will be hard for the contingentist to deny that, from a certain point of view, it *is* possible that I should be s^l . The challenge is not easily dismissed!

3. The Notion of Perspective

The challenge can be met, however, by means of a distinction between two senses of metaphysical possibility. The contingentist may take it to be possible that I should be Fred in one sense of possibility, but concede that in another sense this is not possible. But what is this distinction, and how can this contingentist response be defended?

It might be thought that these questions are readily disposed of by appeal to the framework of centered worlds. A centered world is an ordered pair comprising a possible world and an object said to be its "center." The contingentist might think to proceed in the following way. She may first introduce a notion of truth at a centered *world* in such a way that "I am Fred" will be true at a centered world (w,s) just in case at w, *s* is Fred.[10] She may then define two senses of possibility. For something to be possible in the first sense is for it to be true at some world centered on me. For something to be possible in the second sense, by contrast, is for it to be true at some centered world or other, no matter whom it is centered on.

[10] Cf. Ninan (2009: n. 24).

The contingentist may then offer the following response to the challenge. She may concede that there is no world centered on me at which "I am Fred" is true, and thus that it is not possible in the first sense that I should be Fred. But she may insist that there *is* a world centered on Fred at which "I am Fred" is true and thus that it *is* possible in the second sense that I should be Fred.

In some ways this is not so far from the response we will ultimately recommend. But as it stands it is inadequate. After all, why couldn't someone admit the con- tingentist's domain of centered worlds (for they are nothing but pairs of possible worlds and objects) as well as her notion of truth at a centered world (for one may introduce whatever technical notions one likes) but simply deny that her two definitions correspond to any real forms of possibility? It is not clear why such a position is ruled out. But unless this can be made clear, this response cannot be judged a success.

I believe the desired modal distinction should instead be understood in terms of the notion of a *perspective*.[11] Each of us has a perspective. For example, from my perspective the town of Saxapahaw, North Carolina, is nearby, while from the perspective of Edward Snowden, confined as he is to Moscow, it is far away.

The sense of perspective I have in mind is metaphysical rather than epistemic. Of course, there *is* an epistemic sense of perspective. For example, we may say that from Plato's perspective the soul is tripartite, meaning by this only that Plato takes the soul to be tripartite. But there is also a metaphysical sense of perspective. Even if Snowden has never heard of Saxapahaw and thus does not take it to be any way at all, there is still a sense in which from his perspective it is far away. Saxapahaw's remoteness, that is, is part of the way the world is from his perspective *regardless* of how he takes the world to be.

It is natural to understand a perspective in this sense to itself have both a perspectival and a non-perspectival aspect. Thus not only is it part of how the world is from my perspective that Saxapahaw is nearby, it is also part of how the world is from my perspective that the earth is round and that $2 + 2 = 4$. To be sure, there is also a more restricted notion of perspective which is without any non- perspectival aspect. On this restricted notion, it will be the case from my perspective that Saxapahaw is nearby but not that the earth is round or that $2 + 2 = 4$. But my concern here will be with the unrestricted notion.

It will be helpful to allow ourselves a conception of propositions on which they may be perspectival or non-perspectival.[12] The proposition that Saxapahaw is nearby, for instance, will be perspectival, while the proposition that the earth is round will

[11] Related notions are found in Fine (2005), Hare (2009), and Merlo (2016).

[12] Such a conception is less controversial than it may appear. For one who usually works with a conception of propositions on which they are all non-perspectival may nonetheless be able to recognize a new conception that allows some to be perspectival. One might, for instance, adopt a new conception on which a proposition is taken to be a complex consisting of a proposition in the usual sense together with a mode of presentation. If one thinks there are perspectival modes of presentation, one will arrive at a conception of propositions on which some propositions are perspectival.

not be. Given such a conception we may describe a perspective by specifying which propositions hold from it. Perspectival propositions will serve to describe the perspectival aspect of a given perspective, while non-perspectival propositions will serve to describe its non-perspectival aspect.

Might one take a perspective to have only a non-perspectival aspect, or to be properly described by means of only non-perspectival propositions? On this view it will not be true that from Snowden's perspective Saxapahaw is far away. Instead, it will be true only that from Snowden's perspective Saxapahaw is far away from Snowden. But this strange view cannot accommodate the fact of perspectival difference. Consider Jones, who is visiting North Carolina but has not heard of Saxapahaw. Despite their mutual ignorance of Saxapahaw, Jones and Snowden differ in their metaphysical perspectives on it. But surely no two metaphysical perspectives differ in their non-perspectival aspects. For example, it is equally the case from both perspectives that Saxapahaw is far away from Snowden and is near Jones. This view, then, cannot account for perspectival difference, and so we should not adopt it. We should continue to take a perspective to have both a perspectival and a nonperspectival aspect.

If this metaphysical notion of perspective is admitted, it is very plausible to take it to have, as part of its perspectival aspect, a distinctively first-personal aspect. In the Tractatus, Wittgenstein imagined a complete description of one's perspective in the form of a book titled The World as I Found It.[13] If I were to write such a book, then among its claims would be statements in the first person. For instance, I will write the sentence "I am in North Carolina." Nor is my perspective confined to propositions about the locations of things. For I will also write the sentence "I am MG." And if, for instance, I am in pain, then I will write, "I am in pain." The world, it seems, is somehow given to me in a first-personal way.

Of course, I am not special in this regard. Let us suppose that Snowden writes his own version of The World as I Found It. Then somewhere within its pages we will find the sentences "I am in Moscow" and "I am ES." And if Snowden is in pain, then we will find the sentence "I am in pain." Snowden's perspective, no less than my own, has a first-personal aspect.

To be sure, we can recognize notions of perspective that are in no way first- personal. For instance, there is a sense in which it can be said that from the perspective of an eastbound ship in the Mediterranean, Africa is to starboard. This purely spatial notion of perspective is without any first-personal aspect. But our concern is with the notion of the perspective of a subject, which is plausibly taken to have a first-personal aspect.

If such a notion of perspective is admitted, we face the question of how its first-personal aspect should be described. We might think to describe the perspective of Edward Snowden by saying that from his perspective, Snowden is in Moscow. But although this statement is true, it is not sufficient to describe the first-personal aspect

[13] Wittgenstein (1922: 5.631). Of course, he would not agree with my claims about the book's contents!

of Snowden's perspective. For it is also the case from *my* perspective that Snowden is in Moscow, and so this description fails to capture the first-personal difference between us. In the same way, it is not sufficient to say that from Snowden's perspective he (or he himself) is in Moscow.

Instead, a proper description of the first-personal aspect of Snowden's perspective must itself be first-personal. Indeed, the notion of perspective seems to be an "immersive" one: to describe a perspective we must inhabit it, so to speak, and state how the world is from the resulting standpoint.[14] We should therefore describe the first-personal aspect of Snowden's perspective by means of a proposition that is not merely perspectival but first-personal: from Snowden's perspective, I am in Moscow.

This description cannot be regarded as a piece of ordinary language. After all, the utterance "from Snowden's perspective, I am in Moscow" would ordinarily be taken to describe Snowden's perspective on *my* location, whereas the intent is to describe Snowden's perspective on his own location. I do not wish to deny that there is an ordinary-language sense of perspective in which to say that from Snowden's perspective I am in Moscow is to say something about me. But this notion of perspective appears to be non-immersive, and indeed to lack any first-personal aspect. After all, how in the ordinary sense of perspective are we to capture the perspectival difference between Snowden and me? It is again no help to point out that from Snowden's perspective he (or he himself) is in Moscow and it is not clear how else the difference might be captured. The ordinary notion of perspective, then, is not our notion of perspective, and so we may set it aside.

Since our description of Snowden is not given in ordinary language, we need not worry that it fails to conform to the standard view of the behavior of "I" in ordinary language. Standardly, a token of "I" is held to refer to the agent of the context in which it is tokened.[15] But this standard view does not extend to our description of Snowden, and it is easy to see why. Since our notion of perspective is immersive, a description of a given perspective will be given from the standpoint of the perspective itself rather than the standpoint of the describer of the perspective. And so we should not in general expect a token of "I" in such a description to refer to the describer, despite the describer's being the agent of the relevant context.

This nonstandard behavior creates a risk of confusion which it is wise to guard against. In saying that from Snowden's perspective I am in Moscow, we must remember that the intent is to describe Snowden's perspective on his own location rather than on mine. To remind ourselves we may choose to write "I*" instead of "I" when we are within the scope of a "perspectival operator." For example, we may say that from Snowden's perspective I* am in Moscow while from my perspective I* am in North Carolina. But we should bear in mind that this is only a notational convenience, and that the meaning of the two terms is the same.

[14] Paul's (2016) notion of the "agential perspective" is immersive in this sense.

[15] As in Kaplan (1989).

Even if one takes a proper description of Snowden's perspective to require a first-personal proposition, one might object that there is some ambiguity in our specification of this proposition. We have described Snowden's perspective by saying that from his perspective, I (equivalently: I*) am in Moscow. But one might agree with Frege that " everyone is presented to himself in a special and primitive way in which he is presented to no one else" (1956: 298). And one might think on this basis that a proper description of Snowden's perspective should specify the "special and primitive" sense of "I" in which Snowden is presented to himself. Thus rather than (or in addition to) saying that from Snowden's perspective I am in Moscow, we should say that from Snowden's perspective *IES* am in Moscow, where *"IES"* is understood in this special sense. But even if our description of Snowden's perspective possesses this Fregean ambiguity, little will turn on it, and so for presentational reasons I will usually not bother to resolve it.

Let us admit this notion of perspective. I would then like to suggest that there is reason to accept the following principle:

Veridicality: For any ϕ, ϕ iff from my perspective, ϕ.

Put another way, what is the case is aligned with my perspective.

This formulation of the principle involves quantification into sentence position. We might alternatively formulate the principle without such quantification by saying that for any proposition p, p is true iff from my perspective p is true. (Recall that we admit both perspectival and non-perspectival propositions.)

To see the plausibility of the principle, we need only consider cases. For example, not only is it the case that I am thinking, it is also the case from my perspective that I am thinking. Or again, not only is it the case from my perspective that the moon is closer than the sun, it is also the case that the moon is closer than the sun. Again, not only is it the case that snow is white, it is also the case from my perspective that snow is white. In general, then, it seems that what is the case is aligned with my perspective—that is, that the principle of veridicality holds.

It is important to bear in mind that the notion of perspective involved in the principle is the metaphysical one. The corresponding principle involving the epi- stemic notion would not be plausible. For my epistemic perspective may misrepresent the world: the world may not be the way I take it to be. A metaphysical perspective, by contrast, is not a representation of the world and so there is no possibility of misrepresentation.[16]

All the same, the principle might be challenged. In particular, one might have a concern about its right-to-left direction. For even if ϕ is the case from my perspective, what if there is someone s from whose perspective ϕ is not the case? How then can it be maintained that ϕ is the case? Of course, if the relevant notion of perspective were epistemic there would be no difficulty here: s might simply be mistaken. But if there

[16] I must leave for another time the question of how precisely the relationship between the epistemic

are two metaphysical perspectives that disagree over ϕ, then how can it be the case that ϕ?

I cannot defend any particular answer to this question here. I only wish to insist that there must be some answer.[17] It is implausible that claims over which there is metaphysical-perspectival disagreement should be rejected on that basis. Imagine, for example, that we discover intelligent beings who inhabit a planet in orbit around Proxima Centauri. Although from my perspective the sun is closer than Proxima Centauri, from the perspective of one of these beings the reverse is true. But the discovery of such beings would hardly show that the sun is not, after all, closer than Proxima Centauri.

This point can be strengthened. The special theory of relativity is naturally taken to entail that the simultaneity of events is a perspectival matter. Suppose that this theory is true and that at midnight Greenwich Mean Time fireworks are set off in both London and Edinburgh in celebration of the new year. From my perspective the fireworks are simultaneous. But suppose further that we discover that technologically advanced extraterrestrials have been surveilling Earth from near-light-speed spacecraft. From the perspective of one of these extraterrestrials, the fireworks may well fail to be simultaneous. There are, then, two metaphysical perspectives that disagree over whether the fireworks are simultaneous. But the discovery of such surveillance would do nothing to show that the fireworks are not, after all, simultaneous. Still less would there be any pressure to jettison ordinary claims about, for example, the lengths of familiar objects or the durations of familiar processes, though given relativity these too are naturally taken to be matters of perspective. How could all these utterly quotidian claims be plausibly rejected?

These cases bring out the strength of our commitment to the right-to-left direction of the principle of veridicality. One feels no temptation to abandon it even when confronted with the existence of another perspective that disagrees with one's own.

There is no one else for whom a corresponding principle holds. That is, if t is someone else, then the principle "for any ^, ^ iff from the perspective of t, ^" should not be accepted. After all, the principle of veridicality says that what is the case is aligned with my perspective. The corresponding principle for t can therefore hold only if her perspective completely agrees with my own. But it does not. For at the very least, it is the case from my perspective, but not from that of t, that I am not t (equivalently: that I* am not t).

What is the case, then, is aligned with my perspective and no one else's. We may say on this basis that I am veridical. In this sense, the world is centered on me.

Although I (alone) am veridical, this need not be taken to entail that I am somehow privileged over anyone else. For everyone is veridical from her own perspective. To see this, note first that since I am veridical, the principle of veridi- cality yields the

---

and metaphysical notions should be understood.

[17] Candidates include the external relativism of Fine (2005), the fragmentalism of Fine (2005) and Lipman (2016), the egocentric presentism of Hare (2009), and the subjectivism of Merlo (2016).

conclusion that from my perspective I am veridical (equivalently: I* am veridical). Second, recall that the notion of perspective is immersive. If I "inhabit" the perspective of someone s, I can run through the above reasoning in just the way I have already done, with corresponding results. There is thus reason to say that from the perspective of s I* am veridical.

4. A Defense of Contingentism

The contingentist is now in a position to introduce the modal distinction that is the key to her response to the challenge to her view. To begin, it will be agreed on all sides that there is a sense in which it is impossible that someone else should be veridical. To see this, notice first that it is no accident that the principle of veridicality holds. Think of the cases given above (see Section 2.3) in support of the principle; others could be supplied without limit. They demonstrate a pattern of alignment: something is the case if and only if it is the case from my perspective. It is surely no accident that this pattern obtains.

Consider now someone else s. It is also no accident that what is the case from the perspective of s differs somehow from what is the case from my perspective. After all, every morning I wake up and once again find that I am not s, and so it is no accident that from my perspective I am not s (equivalently: I* am not s). Yet it is also surely no accident that from the perspective of s it is not the case that I* am not s.

Since it is no accident that what is the case is aligned with my perspective, and since it is no accident that s's perspective differs from my own, there is a kind of necessity to the claim that what is the case is not aligned with s's perspective. Put another way, there is a sense in which it is not possible that s should be veridical. Since this is a sense of possibility in which no one else can be veridical, let us call it the proprial sense of possibility (from proprius, "own"). In this sense, the world could not have been centered on someone else.

Although the contingentist should concede that in this proprial sense it is not possible that someone else should be veridical, she may insist that there is also a sense in which this is possible. Her best defense of this claim, I believe, will appeal to the following principle:

> **No one's perspective is impossible:** For any s and ϕ, if from the perspective of s it is the case that ϕ, then it is possible that ϕ.

If there is someone s from whose perspective the world is a certain way, then it cannot be impossible for the world to be that way. After all, from s's perspective it already is that way! She is, if you like, living proof of this possibility.

I would not wish to claim, nor do I believe, that this principle is obviously true. But it has some intuitive appeal. It is certainly not absurd or even implausible. And if the contingentist adopts it, then she has the means for a creditable defense of her position.

The principle, however, must be properly understood if it is not to be rejected out of hand. The sense of possibility involved in the principle cannot be the proprial one. For clearly from the perspective of someone else s, s is veridical, and yet it is not proprially possible that s should be veridical. The principle must therefore be taken to involve some other sense of possibility: a *non-proprial* sense.[18]

We must also bear in mind that the relevant notion of perspective is again metaphysical rather than epistemic. The principle does *not* say that it is always possible for the world to be the way someone takes it to be. It does not entail, for instance, that if someone takes water to be XYZ rather than H_2O then it is possible that water should be XYZ. The principle rather says that if the world is a certain way from someone's perspective *regardless* of how she takes it to be, then it must be possible for the world to be that way.

If we allow the contingentist this principle, then it will be possible that someone else should be veridical. For from the perspective of someone else *s*, *s* is veridical, and so the principle entails that it is possible that *s* should be veridical. In this sense, the world could have been centered on someone else.

This argument must not be misunderstood. The conclusion is not that it is possible that *I* should not be veridical. For no matter who *s* is, it will be true from s's perspective that I* am veridical (equivalently: that I am veridical). Applying the principle to *s* will therefore not entail that it is possible that I should not be veridical; quite the contrary. Instead, the conclusion of the argument is simply that for someone *s* such that I am not *s*, it is possible that *s* should be veridical.

The contingentist may now give her response to the challenge. She may concede that it is not possible in the proprial sense that I should be someone else, such as poor Fred. For it is proprially necessary both that I am veridical and that Fred is not. She may insist, however, that in the non-proprial sense my being Fred *is* possible. After all, it is non-proprially possible that Fred should be veridical, and thus that what is the case should be aligned with Fred's perspective. And surely it is non-proprially necessary that from Fred's perspective, I* am Fred (equivalently: I am Fred). It is therefore non-proprially possible that I should be Fred.

What if the contingentist takes our description of Fred's perspective to possess the Fregean form of ambiguity mentioned above? She will then think that it is the case from Fred's perspective that I_{Fred} am Fred, where *"I_{Fred}"* is understood in the special sense of "I" in which Fred is presented to himself. And she may think there is no other sense of "I" in which from Fred's perspective I am Fred. The fact that from Fred's perspective I_{Fred} am Fred, together with the principle that no one's perspective is impossible, will then entail that it is possible that I_{Fred} should be Fred. But it will not entail that it is possible that I should be Fred in any other sense of "I," including the special sense

[18] I believe we can recognize a temporal counterpart of the present distinction between proprial and non-proprial possibility. The present distinction is therefore more precisely regarded as a distinction between *perspectivally* proprial and non-proprial senses of possibility. I hope to develop the temporal distinction further in future work.

in which MG is presented to himself. The Fregean contingentist may therefore need to qualify her endorsement of contingentism in a way that her non- Fregean counterpart will not.

The contingentist's defense of her view has involved a crucial appeal to the principle that no one's perspective is impossible. Because the principle ranges over all subjects, it yields a strong form of the view. For the strong contingentist, not only is it possible to be someone else, it is possible to be anyone at all, even poor Fred. But one might wish to restrict the range of the principle to certain subjects, the Ss, so as to say only that none of the Ss has an impossible perspective. Such a restricted principle would yield a correspondingly weaker form of contingentism which entails only that it is possible to be one of the Ss.

Indeed, it is only a weaker form of contingentism that is presupposed by those philosophers who think it possible that I should become someone else through a transformative choice. For they need admit only that it is possible to be anyone who might result from my making such a choice, rather than that it is possible to be anyone at all. They may thus defend their view solely by appeal to a restricted version of the principle: no one who might result from my making a transformative choice has an impossible perspective.

5. Conclusions

The contingentist views identity in the way most philosophers have viewed the laws of nature. Such laws are often thought to be necessary in one sense but contingent in another. Thus it is thought that there is a sense in which the energy of an isolated system must remain constant but also a sense in which it is possible for it to increase or decrease. The contingentist thinks the same about who I am. There is a sense in which I must be MG, but there is also a sense in which it is possible that I should be someone else.

This analogy can, I believe, be pursued further. The form of necessity that the laws have has been thought to be somehow less strict than the form of necessity that they lack. Thus although it is necessary that energy is conserved, it is "even more" necessary that two and two are four. No matter what the laws are, after all, two and two will still be four. In a similar way, the contingentist should take the proprial form of necessity, which my identity has, to be less strict than the non-proprial form of necessity, which my identity lacks. Thus although it is necessary that I am MG, it is "even more" necessary that, for instance, everyone is veridical from her own perspective. No matter who I am, this claim about veridicality will still obtain.

Both the contingentist and her opponent, I have argued, should see the world as centered on me. Moreover, both should see this centering as an immutable feature of reality. Yet for the contingentist it is, for all that, not so immutable. There is a sense in which someone else might be found at the center of the world instead. Although I

have the perspective of MG, and have it necessarily, it is nonetheless possible that I should have the perspective of another.

References

Fine, K. (2005). "Tense and Reality." In K. Fine (ed.), *Modality and Tense: Philosophical Papers.* Oxford: Oxford University Press.

Frege, G. (1956). "The Thought: A Logical Inquiry." *Mind* 65(259): 289-311.

Hare, C. (2009). *On Myself, and Other, Less Important Subjects.* Princeton, NJ: Princeton University Press.

Johnston, M. (2010). *Surviving Death.* Princeton, NJ: Princeton University Press.

Kaplan, D. (1989). "Demonstratives." In J. Almog, J. Perry, and H. Wettstein (eds), *Themes from Kaplan.* Oxford: Oxford University Press.

Lewis, D. (1986). *On the Plurality of Worlds.* Oxford: Blackwell.

Lipman, M. A. (2016). "Perspectival Variance and Worldly Fragmentation." *Australasian Journal of Philosophy* 94(1): 42-57.

Merlo, G. (2016). "Subjectivism and the Mental." *Dialectica* 70(3): 311-42.

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

Ninan, D. (2009). "Persistence and the First-Person Perspective." *Philosophical Review* 118(4): 425-64.

Paul, L. A. (2015). "Transformative Choice: Discussion and Replies." *Res Philosophica* 92(2): 473-545.

Paul, L. A. (2016). "The Subjectively Enduring Self." In I. Phillips (ed.), *The Routledge Handbook of the Philosophy of Temporal Experience.* Abingdon: Routledge.

Williams, B. (1973). "Imagination and the Self." In *Problems of the Self: Philosophical Papers 1956-1972.* Cambridge: Cambridge University Press.

Wittgenstein, L. (1922). *Tractatus Logico-Philosophicus.* New York: Harcourt, Brace.

3. How Personal Theories of the Self Shape Beliefs about Personal Continuity and Transformative Experience[(5)]

Sarah Molouki, Stephanie Y. Chen, Oleg Urminsky, and Daniel M. Bartels

1. Introduction

Work on transformative experiences, including some work in the current volume, has tended to focus on normative questions. For example, given the difficulty of grasping the subjective value and experience of living as the transformed individual prior to experiencing a transformation, how can a person make rational decisions related to such experiences? In this chapter, we take a step back and explore people's beliefs about what constitutes a transformative experience. Theoretical characterizations of what generally constitutes transformative experience have been proposed ("you can't know what it is going to be like to be you after the experience. It ... changes your core preferences about what matters": Paul 2014: 17), as well as several canonical illustrative examples (e.g. having a child, becoming a vampire, undergoing a religious conversion, seeing color for the first time; Paul 2014). In contrast, we instead aim to explore the empirical question of what types of changes people believe will transform them into a different individual.

Throughout life, each of us goes through myriad personal changes. However, not all personal changes will transform us into a different individual. For example, a student with a lifelong passion to become a musician will likely undergo significant improvements in knowledge and experience while going through music school. However, we would generally view this personal change to be consistent with, rather than disrup-

[(5)] Sarah Molouki, Stephanie Y. Chen, Oleg Urminsky, and Daniel M. Bartels, How Personal Theories of the Self Shape Beliefs about Personal Continuity and Transformative Experience In: Becoming Someone New: Essays on Transformative Experience, Choice, and Change. Edited by: Enoch Lambert and John Schwenkler, Oxford University Press (2020).

© Sarah Molouki, Stephanie Y. Chen, Oleg Urminsky, and Daniel M. Bartels.

DOI: 10.1093/oso/9780198823735.003.0004

tive of, who she is as an individual. In contrast, most people would consider a sudden loss of this individual's musical abilities due to brain damage to be quite disruptive to her core identity as a musician. When it comes to one's own self-concept, there is considerable individual variation in whether people believe the same change will make them into a different person. For example, one person may think that changing his occupation from accountant to musician will make him into a different person, whereas another person might perceive the same change as preserving rather than disrupting his self-concept. Why do people consider some personal changes to be transformative and others not, and what determines this distinction?

When considering people's subjective beliefs about personal transformation, it may be useful to think in terms of the perceived degree of transformation (e.g. the degree of disruption to the self-concept), consistent with Parfit's idea of varying degrees of psychological connectedness between present and future selves (Parfit 1971; 1984). Given the frequency of personal change that people experience throughout their lives (both positive and negative, as well as expected and unexpected), people may rarely feel as if their self-concept is completely static. By the same token, even dramatic transformative change seldom erases all traces of the pre-existing self. Thus, people's beliefs about personal transformation and disruption largely occur on a continuum of perceived change.

In this chapter, we propose that people have theories about their self-concept that guide which changes and experiences they believe will be more disruptive to who they are and will transform them, to some degree, into a different person. We discuss two different classes of theories that people hold about their self-concepts. First, people have beliefs about the structure of their self-concept—how the different personal aspects are causally related to each other. Second, people have theories about the ways in which their personal qualities will change in the future—how they will develop into the person they expect to become. We propose that personal changes and experiences that are more inconsistent with either of these two types of theories are perceived as more disruptive to the self-concept, and thus more transformative.

Regardless of the normative question of whether people can think rationally about transformative experiences (Paul 2014; 2015), the fact remains that people must and do make decisions about such experiences. Furthermore, recent evidence suggests that understanding whether a person views a particular experience as transformative can be useful for understanding what decisions that person will make. For example, the belief that one will be a different person in the future has an important influence on intertemporal choices. Across various types of decisions, people are more likely to favor their short-term needs over long-term needs when they feel that their future self will be significantly changed from who they are now (see Urminsky 2017 for a review). Thus, understanding people's intuitive theories about transformative personal change may be broadly useful, as these beliefs can affect not only decisions related to the change itself (e.g. whether or not to undergo it) but also broader decisions related to temporal tradeoffs.

2. The Self-Concept and Personal Change

People's self-concept plays a central role in guiding cognition and behaviors in both personal and social domains. In personal decision-making, perceptions of one's own abilities shape goal-setting and motivation (Markus and Nurius 1986; Markus and Wurf 1987); salient aspects of the self can influence current preferences and consumption choices (LeBoeuf etal. 2010; Reed 2004; Reed etal. 2012); and a sense of self-continuity over time can provide the impetus for choosing behaviors with positive long-term consequences when they conflict with short-term desires (Bartels and Rips 2010; Bartels and Urminsky 2011; 2015). In social contexts, the self-concept can serve as a starting point for evaluations of others (Markus and Wurf 1987; Ross et al. 1977) and guide norms and strategies of interpersonal interaction (Akerlof and Kranton 2000; Markus and Wurf 1987; Swann, 1983).

A central question in the study of beliefs about personal transformation is what people believe defines their self-concept. That is, what features, if changed, would make someone into a different person? Many approaches to this question assume that there is a type of feature that tends to consistently be most defining of the selfconcept. For example, some research suggests that mental (rather than physical) changes are the most disruptive to perceptions of personal continuity in other people (Nichols and Bruno 2010). Different researchers have proposed different types of features as particularly defining of the self, including autobiographical memories (Blok etal. 2005; Nichols and Bruno 2010), moral characteristics (Heiphetz etal. 2017; Strohminger and Nichols 2014; 2015), social categories (Akerlof and Kranton 2010; Hogg etal. 1995), and personality traits and preferences (Haslam etal. 2004; Gelman et al. 2007).

According to such feature-based approaches to self, the class of features that is most defining of the self-concept must remain stable to preserve personal continuity. For example, Strohminger and Nichols (2014) suggest that the persistence of one's moral characteristics is most relevant to maintaining a stable self-concept, whereas changes in personal characteristics unrelated to morality will cause relatively little disruption.

An alternative view (Parfit 1984) defines personal continuity in terms of the similarity or degree of overlap in the total set of a person's psychological properties across a particular time period, rather than in terms of the degree of change in a particular type of feature. From this perspective, the greater the magnitude of change (i.e. the greater the overall change in one's total set of psychological features), the greater the disruption to personal continuity. In this view, because there is no core "self" that exists beyond a particular set of psychological features, a larger magnitude of change to these features will disrupt the total constellation of psychological characteristics that defined one's self-concept. Thus, changes to different features may cause different levels of disruption among different individuals based on the degree to which a given feature initially served the purpose of distinguishing that individual from other people.

However, there is reason to believe that neither the type of feature changed nor the total magnitude of the change are the only determinants of what makes an experience disruptive to one's personal continuity. For example, one body of research has found that the valence of change affects perceived personal continuity into the future, with negative changes seen as more disruptive to identity than positive ones (Newman et al. 2014; Newman et al. 2015; Tobia 2015). If a person goes from being extremely cruel to extremely kind, that is viewed as less disruptive to who she is than the same change in the opposite direction (extremely kind to extremely cruel; Tobia 2015). Indeed, people tend to endorse the idea that an individual whose personality and behavior changed from cruel to kind has revealed an aspect of the "true self "—in other words, the positive qualities had always formed a fundamental, potentially hidden, part of her nature (Newman etal. 2014; Tobia 2015), and thus this type of change does not pose any disruption to who she truly is. These findings challenge the idea that feature type or magnitude of change are the only factors that determine personal continuity, since, in this work, feature type and magnitude of change are held constant while only the direction of change is varied.

In the current chapter, we examine recent research on people's intuitive theories about the self and how these theories guide judgments about what changes will transform them into a different person. Rather than focusing on the magnitude of change, the immutability of features, or transient differences in relative salience of features, we examine how the change relates to people's deeper beliefs about their own self-concept, incorporating both past development and expectations about the future. Since these theories of the self vary across individuals, we propose that changes of the same magnitude to a given feature may be reliably perceived as disruptive to some people and as not disruptive to others.

3. How Subjective Theories of the Self Underlie Anticipated Disruption

Lay or intuitive theories are an important part of cognition, influencing knowledge organization, inference, social interactions, and learning. Some researchers suggest that much of cognitive development can be characterized as revision of intuitive theories (Carey 1985; Gopnik and Wellman 1994; 2012; Wellman and Gelman 1992). A classic example of theory revision is the development of theory of mind during which children go from holding a theory that human behavior is driven by the true state of the world to a theory that suggests that behavior is driven by unique internal mental states like desires and beliefs. Thus, while 3-year-olds generally believe that a person will look for a hidden object where it actually is, 5-year-olds have revised these theories and generally believe that a person will look for a hidden object where she thinks it is, regardless of whether the object is actually there (Gopnik and Wellman 1994; but see Baillargeon

etal. 2010 and Rhodes and Brandone 2014 for different perspectives). While there is variation on what is meant by an intuitive theory and its level of specificity, in general, intuitive theories refer to our everyday understanding and explanatory beliefs about the world that often contain beliefs about causal relationships (Carey 1985; Murphy and Medin 1985).

Much of the work to date on intuitive theories has been conducted in the domain of conceptual knowledge, and has examined what types of changes people perceive as disrupting an item's category membership. For example, based on their intuitive theories about immunology, people encountering someone with a new illness they had never heard of before may think that changes to the patients' symptoms (fever and chills) are less likely to change her diagnosis (categorization) than changes to the virus.

In this chapter, we explore personal identity by building on two types of theories identified in the conceptual knowledge literature. First, people have theories about how the features of a concept are causally related to each other, which may include the course of development of those features in the past. These theories allow people to pick out the features that are important to a concept. Features that participate in many cause–effect relationships are generally perceived as more defining of a concept, and changing these features is therefore likely to disrupt category membership (Ahn etal. 2000; Rehder and Hastie 2001). Second, people have domain-specific expectations about how things will develop or change in the future (e.g. an iceberg will shrink but a plant will grow; see Blok etal. 2005; Rips 2011; Rips etal. 2006). Changes that run counter to these expectations are perceived as disruptive to continuity of the object or concept.

Whereas these theories have been extremely influential in the study of biological, artifact, and artificial categories, they only recently been examined in the domain of personal identity (see Chen, Urminsky, and Bartels, 2016 and Molouki and Bartels, 2017). Building on this recent research, we suggest that people have theories about their self-concepts (similar to those they hold about categories) that guide their judgments about the degree to which changes will transform them into a different person. More specifically, people have theories about 1) the causal relationships that exist between the features of their self-concept and, 2) the trajectory of expected or desired change in those features. Changes that are inconsistent with these lay theories—changes that 1) disrupt more of these causal relationships or, 2) deviate from the expected trajectory—will be seen as relatively more disruptive to the selfconcept (i.e. as more transformative) than equivalent changes that are consistent with these theories.

If people's sense of transformation is rooted in their intuitive theories, then it may be more relevant to ask for whom a given change will be seen as transformative, rather than attempt to identify which changes are more uniformly transformative. To the degree that different people have different theories about the causal structure of their self-concepts and different beliefs about their expected trajectory of development, this approach allows us to explain why a given change may be more disruptive for some people than for others.

In the next two sections, we provide theoretical background and review experimental evidence that people's intuitive theories about causal relations and expected change guide their beliefs about what will transform them into a different person. We conclude the chapter with some implications of these beliefs for decisions and behavior, and discuss connections to normative theories of transformative experience.

4. Causal Theories about the Current Self-Concept

We have characterized transformative changes that disrupt personal continuity as changes after which the resulting individual no longer considers him or herself to be an instance of the original individual to a large degree—the original individual has been effectively transformed into a new person. Harry is nearly no longer Harry after a transformative change, but Harry is still largely Harry after a less significant change. These identity judgments, while continuous, are similar to categorization judgments. For example, a duck is no longer a duck when we change a defining feature (its DNA), but it remains a duck when we change a less significant feature (where it lives). From this perspective, the kinds of intuitive theories that guide beliefs about what features define a category and what changes will affect an item's status as a category member may also guide beliefs about what changes people believe will make them a different person, when applied to the self-concept.

Similar to feature and similarity-based approaches to defining personal identity (e.g. Blok et al. 2005; Gelman et al. 2007; Haslam et al. 2004; Nichols and Bruno 2010; Parfit 1984; Strohminger and Nichols 2014; 2015), research about category membership has also explored feature type and magnitude of similarity as key determinants. Early ideas of what features define a category included proposals of specific feature types—e.g. the features that are most frequent among category members (Rosch and Mervis 1975); for example, shape, in the context of artifact categories (Landau etal. 1988). Additionally, many models of categorization suggest that items belong to a concept to the extent that they are similar to other items in the category or to a summary representation (prototype) of the concept, with similarity often defined as the number of features that overlap (Medin and Shaffer 1978; Smith and Medin 1981).

However, it is quite clear that features differ in how central they are to the categorization of an item (Murphy and Medin 1985). For example, the feature *has a fever* seems less defining of the concept of having the flu than the feature *has the influenza virus*. That is, we're more likely to think a person who does not have the virus (but has a fever and chills) does not have the flu than someone who does not have a fever (but has the virus and the chills). This belief tends to hold even though the magnitude of similarity to a prototypical flu patient is the same in both cases (i.e. both have two out of three of the classic flu features).

Theory-based approaches can address these shortcomings of similarity-based models of categorization. In this view, mental representations of concepts are not simply

feature lists, but include theories about how these features are causally related to one another (see Fig. 3.1). These causal relations guide how defining features are of a concept (Murphy and Medin 1985; Ahn et al. 2000). A seminal finding from this work is that a feature is seen as defining a concept to the extent that it is *causally central* to the concept (Ahn 1999; Ahn et al. 2000; Sloman et al. 1998; Rehder and Hastie 2001).

Although there are competing perspectives on the definition of causal centrality, in this chapter it is defined as the number of causal relationships that a feature participates in (either as cause or effect) with other features of the concept (see Chen et al. 2016, experiment 3, for a test of different approaches to causal centrality in the context of the self-concept). For the flu example above, *has the influenza virus* is more defining of the concept of the flu than *has a fever* because the virus is causally linked to two other features (causes both the fever and chills), while *has a fever* is only causally linked to one other feature (it is caused by the virus). Several research streams in cognitive psychology have demonstrated the role of causal knowledge in conceptual centrality. In particular, causal relations between features are often more important for determining category membership than the nature of the features themselves (e.g. Ahn etal. 2000; Rehder and Hastie 2001).

In this section, we examine the proposal that people's subjective representations of the self-concept likewise include causal relationships, and that these causal beliefs influence which aspects of the self are seen as most defining of the self-concept. For example, imagine two people, Jack and Mack. Both have memories of being lonely children, prefer solitary activities, and are very shy. Jack and Mack both believe that these three features are important to their self-concept; however, they differ in how they believe these features are causally related to one another (see Figure 3.1). Jack believes that it was his memories of being a lonely child that caused him to develop into a shy person and prefer solitary activities. In contrast, Mack instead believes that it was his shyness that caused him to be a lonely child and prefer solitary activities.

Previous accounts of the self-concept that assume that a particular type of feature (e.g. memories) is most central would assume that since both Jack and Mack have the same features, they would view changes the same way. However, a causal centrality account of the self predicts that even if the features of the two men are identical, their self-concepts will be fundamentally different because of their differing beliefs about the causal relationships between these features. That is, shyness will be more central to Mack's self-concept than Jack's because it is causally linked to both his preferences and his memories (whereas for Jack it is only causally linked to his memories). As a result, they are likely to view changes differently. In particular, if Mack became outgoing (a change to his shyness, a causally central feature for him) he would feel more like a different person than Jack would after the same change (a change to a less central feature for him).

To test this hypothesis, Chen etal. (2016, supplemental experiment 4, appendix A1) manipulated the same features to be either causally central or peripheral in a series of vignettes that described hypothetical characters' intuitive theories of how the features

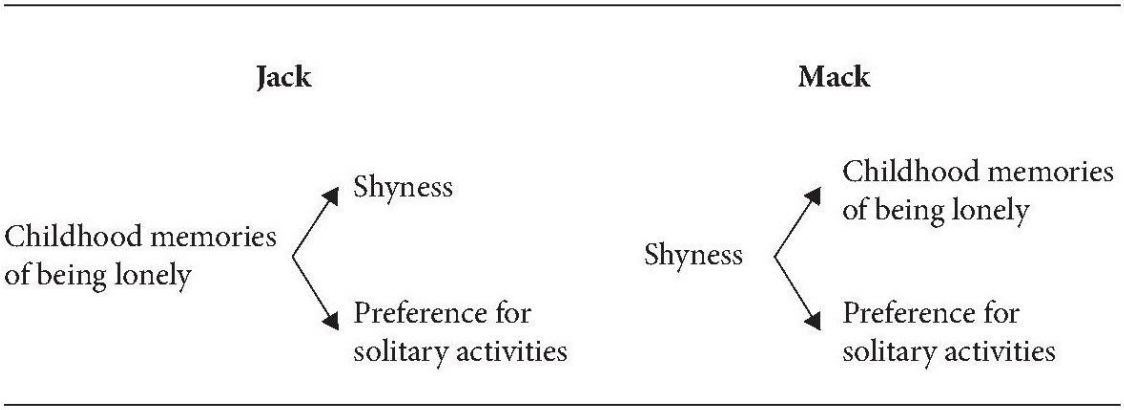


Figure 3.1 Example of causal structure used in Chen et al. (2016: supplemental experiment 4) vignettes. In the Jack vignette, memories are causally central and personality (shyness) is causally peripheral. In the Mack vignette, memories are causally peripheral and personality (shyness) is causally central.

of their self-concepts fit together (like the ones described above for Jack and Mack). Since, in the vignettes, the cause feature had more causal connections than the effect feature, it was relatively more causally central. Each vignette had two versions (like the Jack and Mack versions), manipulating the causal centrality of two focal features by switching which of the two features was a cause and which was an effect (see Figure 3.1). So, the exact same features were counterbalanced to play either the cause or effect role, to control for any idiosyncratic influences of specific features. Subjects only read one version of each vignette—e.g. half the participants read about Jack and half read about Mack.

After reading a vignette about a single person, subjects then chose which of two post-change individuals was more likely to have been the character in the vignette. Each individual was missing one of the manipulated features (e.g. memories or shyness) but retained the other two features. If the causal centrality of a feature influences continuity judgments, subjects should be more likely to pick the individual who retains the causally central feature and is missing the causally peripheral feature (e.g. the person missing shyness if they read the Jack version, and the person missing memories if they read the Mack version).

Consistent with the causal centrality view of the self-concept, on average, participants indicated that the person retaining the causally central feature was the individual initially described in the vignette 68% of the time. This finding suggests that changes in a feature were perceived as more transformative (i.e. Jack seems more like a different person) when that feature was manipulated to be causally central, based on its relationships to the other features as specified in the intuitive theory (vignette), than when the exact same feature was manipulated to be causally peripheral. Additionally,

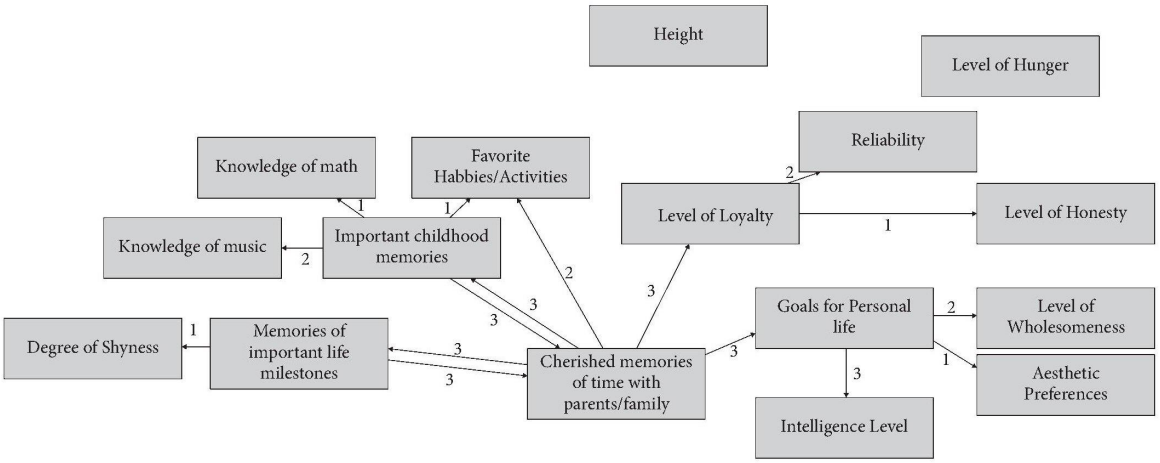
these results cannot be explained by the magnitude of the changes (defined by the amount of feature overlap between the original and changed individual) because the magnitude of change from the original character was the same for both individuals (i.e. only one feature changed and both retained the two other features of the original character).

In the previously described experiment, causal role (whether a feature was a cause or an effect) and causal centrality (number of relationships with other features) were confounded. As the cause feature was always linked to relatively more features than the effect feature, a possible alternative explanation for the previous results is that causes are perceived as more defining of the self-concept than their effects.[1] In a follow-up vignette study, Chen etal. (2016, study 3), examined whether it was the causal role or the number of causal links that drove identity judgments. The results of this study revealed that the number of causal links a feature participated in had greater influence on identity judgments than did the feature's causal role. Changes to the causally central feature were perceived as more disruptive to the self-concept both when the cause feature was described as having relatively more causal relationships (replicating the results of the previous experiment) and also when the effect feature was described having relatively more causal relationships. For example, if the effect feature, Jack's shyness, had three causes—his preference for solitary activities, his memories of being a lonely child, and his awkward demeanor—then a change to the effect feature (shyness) was more likely to be perceived as disruptive to Jack's personal continuity than a change to a cause feature (memories).

To test whether changes to causally central features are perceived as more disruptive to one's own (rather than another person's) self-concept, Chen etal. (2016, experiment 1) measured people's beliefs about the causal centrality and importance of features of their self-concepts. The features of the self-concept were selected from five categories of features suggested to be important in previous research: memories, moral qualities, personality traits, preferences, and goals and desires. Beliefs about causal relationships were measured by having participants verbally report which features they believed had shaped or influenced the other features of their selfconcept. The causal centrality of each feature was then calculated as the number of direct links to other features (either as cause or effect). Beliefs about how defining features are to the self-concept were measured by asking participants how much a change to each feature would disrupt their personal continuity (i.e. the extent to which a change in a feature would lead them to feel that they were a different person).

---

[1] Some approaches to causal centrality in the concepts and categories literature suggest that causes are more central than their effects (Ahn etal. 2000; Sloman etal. 1998; see Rehder 2003 for discussion of various approaches to causal centrality in categorization). However, across a number of experiments Chen et al. (2016) consistently found that the number of causal links better described how causal beliefs influence identity judgments than causal role did.



The arrow starts at the cause feature and points to the effect feature. The numbers that are on each arrow (i = weak, 2 = moderate, 3 = strong).

Consistent with the results of the previously described experiment, people perceived changes to causally central features of their self-concepts as more transformative than changes to causally peripheral features. The number of causal relationships that a feature was seen as participating in was correlated with how disruptive the person perceived a change to that feature to be. This positive relationship between causal relationships and disruption to the self-concept was observed for the large majority of participants (77%). These results were replicated when using an alternative method for eliciting causal centrality adapted from Sloman etal. (1998), the concept map task (Chen et al. 2016: supplemental experiment, appendix A1). In this task, participants drew these causal links in a concept map (see Figure 3.2).

The studies described above demonstrate that people's intuitive theories about how the features of their self-concept are causally interrelated influence their beliefs about which changes will transform them into a different individual. Changes to features that are perceived as causally central—participating in many cause–effect relationships with other features—are perceived as more disruptive to the selfconcept than changes to features with fewer cause–effect relationships. This approach explains both why some features tend to be more defining of the selfconcept than others, on average (i.e. because they are viewed as casually central for more people) and also why some features matter more to some people than to others (i.e. because people have different intuitive theories about their self-concepts).

5. Dynamic Theories about the Future Self-Concept

In addition to having theories about how personal features interact to form the selfconcept (which incorporate causal theories of prior development), people also have ideas about their future trajectories. We propose that people's conception of their future continuity is not solely based on the preservation of a specific type of feature or the proportion of features preserved. Instead, people have expectations and desires about how their personal qualities will change in the future. Changes that are in line with these beliefs will be seen as consistent with the self-concept, whereas changes that are contrary to these expectations and desires will disrupt self-continuity.

Our focus in this section is on the trajectory of change, rather than on which particular feature is changing. This approach contrasts with some research (described in Section 2) that has linked perceived personal continuity to the continuity of specific classes of features (e.g. Strohminger and Nichols 2014). Instead, we describe how people's beliefs about the form or direction of expected change affect perceptions of continuity, across a wide variety of feature types.

One major expectation that people hold about the trajectory of their future change concerns the valence of change—i.e. whether they will improve or decline over time. As described in Section 2, people perceive positive change to be more consistent with their ideas about general human development (Newman et al. 2014; Newman et al. 2015; Tobia 2015). Furthermore, when it comes to ideas about their own personal development, people are even more likely to endorse ideas that they will improve in the future. For example, people state that their own improvement on various personality characteristics will be larger than those of their peers (Kanten and Teigen 2008), and there is recent evidence that people's predictions of their own improvement are in fact over-optimistic compared to their actual change over the same time period on various measures of personality, values, and cognitive performance. This imbalance occurs because people tend to anticipate positive future change but neglect the possibility of negative change in the future (Molouki etal. 2017). Because people expect an improving future trajectory, it is likely that the idea of broad positive change is incorporated into their self-concept. In line with this idea, we propose that people would consider many types of positive changes in themselves (regardless of the type of feature changing) to be identity-consistent, whereas negative changes would be seen as disruptive to personal continuity.

5.1 The Joint Influence of Feature Type and Valence of Change in Perceived Personal Transformation

Molouki and Bartels (2017) tested the theory that perceptions of one's own personal continuity are determined by congruence with a trajectory of improvement across var-

ious types of mental features. Participants were asked to imagine that each of a list of personal characteristics (presented in Table 3.1) would either improve, worsen, or change (valence unspecified) in the future. For each characteristic, they provided their perception of the extent to which the specified change would lead to a discontinuity or disruption in their self-concept. If people incorporate an overall trajectory of improvement into their self-concept, then positive change would be viewed as less disruptive than negative change across all types of features.

In terms of feature type, it was found that changes in items related to morality were perceived as the most disruptive to self-continuity, followed in turn by changes in items related to personality, preferences, experiences, and memories (see Figure 3.3). These results for people's beliefs about their own change are generally consistent with findings demonstrating an influence of the type of feature changing on judgments of the continuity of third parties (Strohminger and Nichols 2014).

However, importantly, the degree of perceived disruption to personal continuity also depended on the valence of the change. On average, positive change was perceived to be less disruptive to the self-concept than either negative change or unspecified change for all categories of characteristics, suggesting that beliefs about personal continuity are not solely guided by the type of feature changing. Thus, although changes to morality and personality generally lead to the greatest perceptions of discontinuity, change that was explicitly labeled as positive was an exception, such that disruption to the self-concept from positive change was quite low for all types of features (Figure 3.3).

Table 3.1 Listing of stimuli used to measure the effects of positive and negative change across various feature types. Characteristics were selected and categorized based on pre-tests (Molouki and Bartels 2017).

Morality	***Personality***
Morality	
Values	
Degree of honesty	
Beliefs	
Level of humility	
Religious or spiritual faith	
Level of selflessness	Sense of humor
Impulsiveness	
Temperament	
Curiosity	
Dominance	
Level of calmness	
Confidence level	
Imagination	
Reliability	
Degree of independence	
Level of friendliness	
Self-awareness	
Level of happiness	
Intelligence level	
Level of disorganization	
Helpfulness	

Preferences	***Experiences***	***Memories***
Professional goals		
Preferences/Favorite things		
Major likes and dislikes		
Goals for your personal life	Life	
experiences		
Friendships		
Everyday activities		
Health		
Ability to feel pain		
Occupation	Cherished memories of time Spent with family	
Memories of time spent Commuting to work		
Knowledge of how to ride a bike		
Knowledge of how to play the piano		
Knowledge of math		
Bad memories		

These results suggest that the valence of change influences whether change is seen as disruptive to self-continuity. Beliefs about improvement being consistent with the self may be common to most individuals, as they may stem from intuitive beliefs about positive human development over time (Newman et al. 2014; Tobia 2015), or be related to common aspirations for self-improvement (Kanten and Teigen 2008). However, the impact of valence on perceived personal continuity may be further defined by a person's specific expectations about her individual trajectory of personal change—i.e. how she expects her own particular characteristics to change in the future, as defined by her unique view of her own self-concept.

5.2 The Effect of Individual Expectations about Specific Changes on Perceived Personal Transformation

Research on how people categorize objects suggests that beliefs about identity continuity depend on expectations of change, which vary across different types of objects (Blok etal. 2005; Gutheil etal. 2004; Gutheil etal. 2008; Gutheil and Rosengren 1996, Rips 2011; Rips etal. 2006). For example, people believe that the process of growing

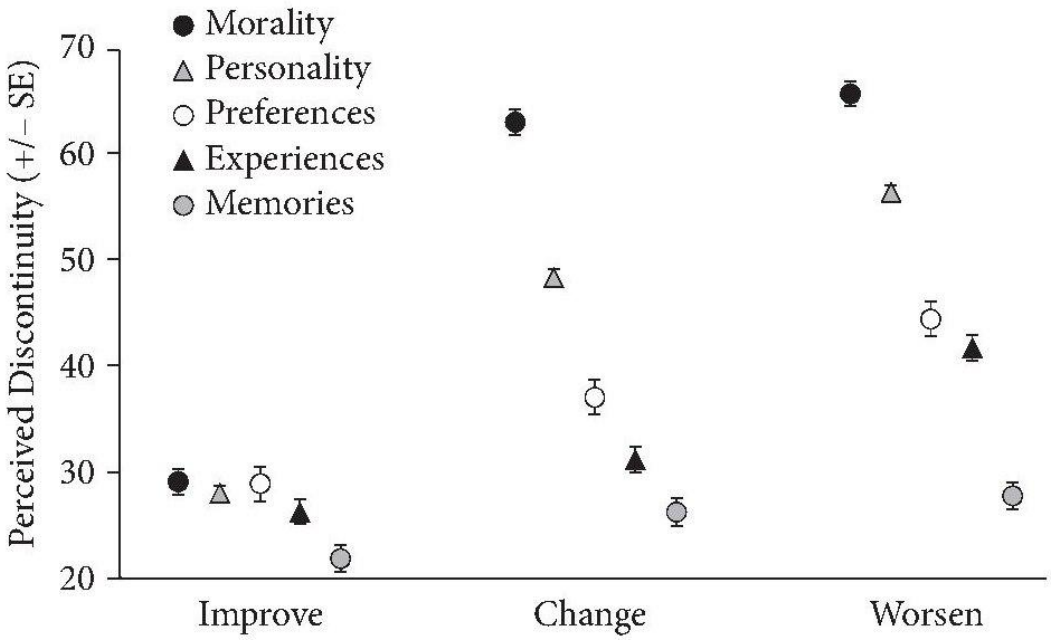


Figure 3.3 Perceived self-discontinuity ratings by valence and category of change. (Figure reproduced with permission from Molouki and Bartels 2017.)

smaller is consistent with the identification of an object as an iceberg (see Rips etal. 2006), whereas the process of growing larger would be consistent with the identification of a different type of object, such as a plant. Observing the opposite pattern (an iceberg growing larger or a plant growing smaller) would be unexpected, and might suggest that the object does not fit into the proposed category or that a disruption to its continuity has occurred.

Comparing change to an individually defined trajectory to determine category membership is a central feature of several theories of categorization of objects in the physical world (Blok et al. 2007; Nozick 1981; Rips 2011; Rips and Hespos 2015; Rips et al. 2006; Sagi and Rips 2014). We propose that much as a person might use specific beliefs about the natural course of development of a certain object to assess whether it has maintained its identity over time (e.g. an iceberg will shrink over time, but a plant will grow over time), she might also contrast a particular change with her expectations about her own natural development (e.g. idiosyncratically determined personal expectations and aspirations) to assess whether this change would disrupt or maintain her personal continuity over time.

Intuitive theories of positive personal development could help explain why positive change in general tends not to lead to disruption of personal continuity, since positive change tends to be consistent with most people's expectations and desires (e.g. Bench etal. 2015; Busseri etal. 2009; Haslam etal. 2007; Molouki etal. 2017; Newby-Clark and Ross 2003; Wilson and Ross 2001). Thus, whereas people may have already incorporated the idea of positive change into not only their planned life trajectory but also their self-concept, negative change is more likely to be inconsistent with both expectations and desires for their personal development over time, and therefore more disruptive. However, specific expectations about a given characteristic might also guide the perceived effects of change in this particular feature on self-continuity.

To test this idea, Molouki and Bartels (2017) orthogonally manipulated both valence of change and expectations of change to isolate the influence of these factors. This was achieved by asking participants to select features for which they held strong expectations about change, and then asking them to imagine that each of these features in turn would improve, worsen, or stay the same. If individual expectations play a role in judgments of the self-concept, any type of change, regardless of valence, that is in alignment with a person's expectations for her own future will result in relatively less disruption to perceived personal continuity than a change that is misaligned with expectations.

Molouki and Bartels (2017) found that people's individual expectations indeed influenced their perceptions of continuity. Although people's perceptions of disruption were influenced by the overall valence of change (improvements were seen as less disruptive), changes that were consistent with a participant's specific expectations about future change led to relatively lower perceived disruption to personal continuity (see circled means in Figure 3.4). In contrast, change in any characteristic that ran counter to one's theory of personal change was associated with increased reports of disruption.

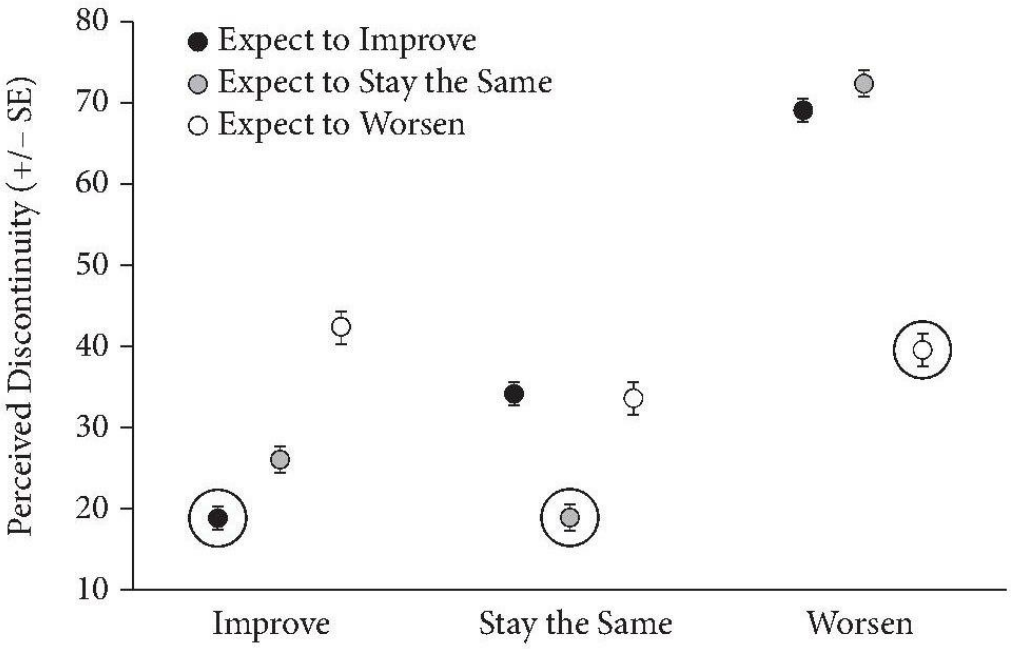


Figure 3.4 Perceived self-discontinuity ratings by imagined change and expected change. Note: Circled values are those for which imagined change matches expected change. (Figure reproduced with permission from Molouki and Bartels 2017.)

Decline was viewed as relatively more disruptive to the selfconcept when a person specifically expected improvement in that characteristic, whereas improvement was seen as relatively more disruptive when a person specifically expected decline in that characteristic.

5.3 The Effect of Individual Desires for Specific Changes on Perceived Personal Transformation

People tend to expect improvements for their future change, but it is not clear whether these ideas about future improvement stem from specific personal desires, or are instead guided by broader lay theories about how people, in general, change over time (e.g. Newman etal. 2014). This is an important distinction because if people evaluate whether a change is disruptive to their personal continuity by consulting their own individual desires, it suggests that people's expectations about future development, and, in turn, beliefs about the self-concept, may contain a specific aspirational component that varies from individual to individual based on their personal goals.

Molouki and Bartels (2017) found that congruence of a change with people's specific desires mattered for continuity judgments. Changes that were consistent with specific desires for change in a given feature (see circled means in Figure 3.5) led to less perceived discontinuity than undesired changes. This suggests that changes that were specifically undesired were more likely to be judged as disruptive to the selfconcept. The effect of desires on continuity judgments was distinct from the main effect of valence. In other words, although people reported less discontinuity on average in response to improvements than in response to worsening in any characteristic (a general valence effect), a given participant was also more likely to report feeling that her personal continuity was disrupted when her specific desires for a given characteristic were not met (an effect of individual desires).

The results of this study suggest that individual desires play a significant role in perceptions of self-continuity, above and beyond the general effect of valence of change. For example, rather than indiscriminately reporting that improvements would be the most consistent with their self-concept, people actually reported somewhat greater discontinuity when a trait they desired to stay the same improved, compared to when it stayed the same. Overall, the effect of specific desires on selfcontinuity judgments reveals that people incorporate their idiosyncratic personal desires and goals into theories about the development of their self-concept.

6. Conclusions

In this chapter, we have summarized recent empirical findings which suggest that people's intuitive theories about their present and future self-concept determine whether a change is more likely to be perceived as transforming them into a different

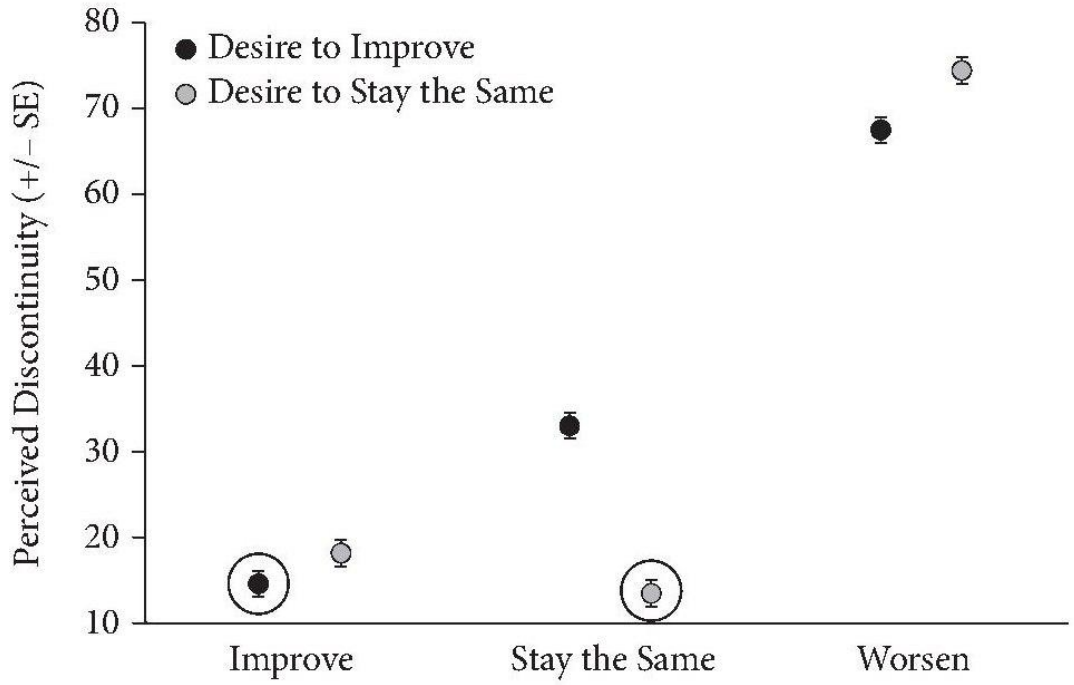


Figure 3.5 Perceived self-discontinuity ratings by type of change and desire for change. Features with desire for worsening were not identified in this experiment, as it was expected that most people would not hold such desires. Note: Circled values are those where imagined change matches desired change. (Figure reproduced with permission from Molouki and Bartels 2017.)

individual. Based on these findings, the most transformative changes are likely to be those which are unexpected, unwelcome, and involve aspects of the self that are most causally linked to other features. However, these are not mutually necessary conditions. For example, even a positive, desirable change can be seen as transformative (i.e. as causing a discontinuity in the self) when the change is highly unexpected. Our framework also predicts that unexpected and unwelcome changes to less causally central features will be less disruptive, and that welcomed and perhaps even expected changes may nevertheless be seen as transformative if the change is to a highly causally central aspect of the self-concept. However, these predictions about the interaction between developmental expectations and causal centrality have not yet been empirically tested.

The concept of personal disruption or discontinuity that we have introduced in this chapter provides a new way of understanding how people think about transformative experiences. Although the phrase "transformative experience" has previously been used to refer to a characteristic of specific life events (e.g. having a child, experiencing a religious conversion, becoming a vampire), the defining nature of such experiences is that they involve a change, sometimes unexpected or unwelcome a priori, to fundamental preferences, beliefs, or ways of seeing the world (Paul 2014). Our experiments evaluated how people's intuitive theories relate to their beliefs about the impact of various types of changes in these features (preferences, beliefs, and other personal characteristics). In this view, a variety of life experiences can be subjectively perceived as transformative to varying degrees, depending on people's beliefs about changes to these features.

In many cases, these individual perceptions may differ from normative definitions of what constitutes a transformative experience. For example, some normative descriptions may classify the birth of a child as a transformative change, since it brings about unanticipated developments to one's fundamental ways of thinking and functioning as a person that could not be fathomed prior to the change (Paul 2014; 2015). However, individuals themselves will likely vary on the degree to which they view having a child as a transformative experience. Though some individuals believe that the birth of a child will transform who they are as a person, others might not anticipate that this event would disrupt who they are because it is viewed as (a) consistent with their causal theories of their self-concepts, (b) a positive change, and/ or (c) an expected or desired change. If a person sees having a biological child in the future as desirable, expected, and causally linked to many other aspects of the self, then it may in fact be the failure to conceive that would be assessed as far more of a disruption to the self-concept.

The work reviewed in this chapter suggests that intuitive theories of the selfconcept, specific to an individual, influence the degree to which a change is seen as transformative. Variations across individuals' intuitive theories influence the degree to which an anticipated change makes the person feel that the future self will be continuous with the current self or transformed into a different person. Such beliefs may be an important consideration in decision-making in several different ways.

People may choose to behave in ways that are shaped by the anticipation of disruption to the self-concept. If people are generally motivated to preserve their continuity, they may be more likely to act in ways that are consistent with the causally central features of their identity. Recent research finds that people who perceive that a social category (e.g. Democrat or Republican) is causally central in their self-concept are more likely to act in ways that are consistent with that aspect of their identity (e.g. voting for the candidate nominated by their political party) than those who believe that the same social category is causally peripheral (Chen and Urminsky 2019). Understanding the causal structure of a person's identity may thus provide unique insight into their likely identity-consistent future behavior. In one study, people for whom the feature of honesty was more causally central were less likely to cheat in an incentivized coin-flipping game, while self-reported level of honesty and importance of honesty were not predictive (Chen and Urminsky 2017). However, in other cases, people may not be motivated by identity-consistency. For example, when people anticipate future negative outcomes, they may instead be motivated to act in ways that disrupt their identity, or at least signal a change in identity (Yang and Urminsky 2015).

When change is inevitable, beliefs about transformation may also shape how people cope with that change. For example, Parfit (1984) points out that if one perceives the future self to be only weakly connected to the current self, this may reduce one's current concern about the fate of this future self. Given that a transformative experience will tend to weaken the perceived connection between current and future self, the very act of perceiving an experience as transforming the self could thus mitigate concern and anxiety about the future outcomes of that change. Nichols et al. (2017) discuss a similar concept in the realm of religious beliefs about the self. In a study of participants of various religions, Tibetan Buddhists, who endorse the idea that there is no such thing as a continuous, enduring self, reported using this belief to cope with thoughts of death (perhaps the most extreme transformative experience possible). However, Buddhist participants nevertheless reported high levels of anxiety about self-annihilation at death, suggesting that feelings or intuitions about self-continuity may persist despite explicit religious beliefs. An interesting topic of future research is the relationship between beliefs about personal transformation and strategies for coping with unexpected or undesired changes.

Beliefs about transformation and disruption may have broad implications for understanding people's motivation for future-oriented decision-making (Bartels and Rips 2010). When changes on the horizon are seen as transformative, reducing people's subjective sense of similarity to their future self, they are likely to feel less motivated to sacrifice in the present for the benefit of that future self. This can have important consequences for decision-making, including greater impatience (Bartels and Urminsky 2011), reduced motivation (Peetz etal. 2009; Dai etal. 2015), more willingness to spend discretionary funds even when considering future consequences (Bartels and Urminsky 2015), and the commission of unethical acts (Hershfield et al. 2012). On the other hand, disruption to self-continuity and the resulting reduced concern about the future

self can also prompt positive behaviors such as heightened generosity to future others (Bartels et al. 2013).

Lastly, beliefs about change may shape how people make decisions about the change itself. An individual who does not consider childbirth to be a transformative experience may be confident (or even overconfident) in her abilities to make decisions about this event, believing that she will remain mostly the same person with the same preferences, beliefs, desires, etc. after having a child. In contrast, someone for whom becoming a parent is less consistent with her theories of her own self-concept might not be as confident in making decisions about how to parent. Because she is more likely to think that the experience will transform her into a different individual, she may feel that she does not fully grasp how to evaluate and respond to the outcomes of this decision.

Thus, evaluating a given individual's perceptions about what constitutes a personally transformative experience is an important supplement to research on normative theories of decision-making in this domain, as people's tendencies to employ decision-making strategies will vary based on their perceptions of the nature of the experience. Regardless of whether or not a given individual's theory of transformative experience agrees with or diverges from normative views, it is ultimately the individual's own evaluation that will guide her feelings and behaviors surrounding the experience.

References

Ahn, W. K. 1999. "Effect of Causal Structure on Category Construction." *Memory and Cognition* 27(6): 1008-23.

Ahn, W. K., N. S. Kim, M. E. Lassaline, and M. J. Dennis. 2000. "Causal Status as a Determinant of Feature Centrality." *Cognitive Psychology* 41(4): 361-416.

Akerlof, G. A., and R. E. Kranton. 2000. "Economics and Identity." *Quarterly Journal of Economics* 115(3): 715-53.

Akerlof, G. A., and R. E. Kranton. 2010. *Identity Economics: How Our Identities Shape Our Work, Wages, and Well-Being.* Princeton, NJ: Princeton University Press.

Baillargeon, R., R. M. Scott, and Z. He. 2010. "False-Belief Understanding in Infants." *Trends in Cognitive Sciences* 14(3): 110-18.

Bartels, D. M., T. Kvaran, and S. Nichols. 2013. "Selfless Giving." *Cognition* 129(2): 392-403.

Bartels, D. M., and L. J. Rips. 2010. "Psychological Connectedness and Intertemporal Choice." *Journal of Experimental Psychology: General* 139(1): 49-69.

Bartels, D. M., and O. Urminsky. 2011. "On Intertemporal Selfishness: How the Perceived Instability of Identity Underlies Impatient Consumption." *Journal of Consumer Research* 38(1): 182-98.

Bartels, D. M., and O. Urminsky. 2015. "To Know and to Care: How Awareness and Valuation of the Future Jointly Shape Consumer Spending." *Journal of Consumer Research* 41(6): 1469-85.

Bench, S. W., R. J. Schlegel, W. E. Davis, and M. Vess. 2015. "Thinking About Change in the Self and Others: The Role of Self-Discovery Metaphors and the True Self." *Social Cognition* 33(3): 169-85.

Blok, S., G. Newman, and L. J. Rips. 2005. "Individuals and Their Concepts." In W. K. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, and P. Wolff (eds), *Categorization Inside and Outside the Lab*, 127-49. Washington, DC: American Psychological Association.

Blok, S., G. Newman, and L. J. Rips. 2007. "Out of Sorts? Some Remedies for Theories of Object Concepts: A Reply to Rhemtulla and Xu (2007)." *Psychological Review* 114(4): 1096-102.

Busseri, M. A., B. L. Choma, and S. W. Sadava. 2009. "Functional or Fantasy? Examining the Implications of Subjective Temporal Perspective 'Trajectories' for Life Satisfaction." *Personality and Social Psychology Bulletin* 35(3): 295-308.

Carey, S. 1985. *Conceptual Change in Childhood.* Cambridge, MA: Bradford Books.

Chen, S. Y., and O. Urminsky. 2017. "We Are What We Think: Representations of the SelfConcept and Identity-Based Choice." University of Chicago Working Paper.

Chen, S. Y., and O. Urminsky. 2019. "The Role of Causal Beliefs in Political Identity and Voting." *Cognition* 188: 27-38.

Chen, S. Y., O. Urminsky, and D. M. Bartels. 2016. "Beliefs About the Causal Structure of the Self-Concept Determine Which Changes Disrupt Personal Identity." *Psychological Science* 27(10): 1398-406.

Dai, H., K. L. Milkman, and J. Riis. 2015. "Put Your Imperfections Behind You: Temporal Landmarks Spur Goal Initiation When They Signal New Beginnings." *Psychological Science* 26(12): 1927-36.

Gelman, S. A., G. D. Heyman, and C. H. Legare. 2007. "Developmental Changes in the Coherence of Essentialist Beliefs About Psychological Characteristics." *Child Development* 78(3): 757-74.

Gopnik, A., and H. M. Wellman. 1994. "The 'Theory Theory.' " In L. Hirschfield and S. Gelman (eds), *Domain Specificity in Culture and Cognition,* 257-93. Cambridge: Cambridge University Press.

Gopnik, A., and H. M. Wellman. 2012. "Reconstructing Constructivism: Causal Models, Bayesian Learning Mechanisms, and the Theory Theory." *Psychological Bulletin* 138: 1085-1108.

Gutheil, G., P. Bloom, N. Valderrama, and R. Freedman. 2004. "The Role of Historical Intuitions in Children's and Adults' Naming of Artifacts." *Cognition* 91(1): 23-42.

Gutheil, G., S. A. Gelman, E. Klein, K. Michos, and K. Kelaita. 2008. "Preschoolers' Use of Spatiotemporal History, Appearance, and Proper Name in Determining Individual Identity." *Cognition* 107(1): 366-80.

Gutheil, G., and K. S. Rosengren. 1996. "A Rose by Any Other Name: Preschoolers' Understanding of Individual Identity Across Name and Appearance Changes." *British Journal of Developmental Psychology* 14(4): 477-98.

Haslam, N., B. Bastian, and M. Bissett. 2004. "Essentialist Beliefs About Personality and Their Implications." *Personality and Social Psychology Bulletin* 30(12): 1661-73.

Haslam, N., B. Bastian, C. Fox, and J. Whelan. 2007. "Beliefs About Personality Change and Continuity." *Personality and Individual Differences* 42(8): 1621-31.

Heiphetz, L., N. Strohminger, and L. L. Young. 2017. "The Role of Moral Beliefs, Memories, and Preferences in Representations of Identity." *Cognitive Science* 41(3): 744-67.

Hershfield, H. E., T. R. Cohen, and L. Thompson. 2012. "Short Horizons and Tempting Situations: Lack of Continuity to Our Future Selves Leads to Unethical Decision Making and Behavior." *Organizational Behavior and Human Decision Processes* 117(2): 298-310.

Hogg, M. A., D. J. Terry, and K. M. White. 1995. "A Tale of Two Theories: A Critical Comparison of Identity Theory with Social Identity Theory." *Social Psychology Quarterly* 58(4): 255-69.

Kanten, A. B., and K. H. Teigen. 2008. "Better Than Average and Better With Time: Relative Evaluations of Self and Others in the Past, Present, and Future." *European Journal of Social Psychology* 38(2): 343-53.

Landau, B., L. B. Smith, and S. S. Jones. 1988. "The Importance of Shape in Early Lexical Learning." *Cognitive Development* 3(3): 299-321.

LeBoeuf, R. A., E. Shafir, and J. B. Bayuk. 2010. "The Conflicting Choices of Alternating Selves." *Organizational Behavior and Human Decision Processes* 111(1): 48-61.

Markus, H., and P. Nurius. 1986. "Possible Selves." *American Psychologist* 41(9): 954-69.

Markus, H., and E. Wurf. 1987. "The Dynamic Self-Concept: A Social Psychological Perspective." *Annual Review of Psychology* 38(1): 299-337.

Medin, D. L., and M. M. Schaffer. 1978. "Context Theory of Classification Learning." *Psychological Review* 85(3): 207-38.

Molouki, S., and D. M. Bartels. 2017. "Personal Change and the Continuity of the Self." *Cognitive Psychology* 93: 1-17.

Molouki, S., D. M. Bartels, and O. Urminsky. 2017. "Neglecting Decline: Remembered and Predicted Personal Development Diverge from Actual Longitudinal Change." University of Chicago Working Paper.

Murphy, G. L., and D. L. Medin. 1985. "The Role of Theories in Conceptual Coherence." *Psychological Review 92(3):* 289-316.

Newby-Clark, I. R., and M. Ross. 2003. "Conceiving the Past and Future." *Personality and Social Psychology Bulletin* 29(7): 807-18.

Newman, G. E., P. Bloom, and J. Knobe. 2014. "Value Judgments and the True Self." *Personality and Social Psychology Bulletin* 40(2): 203-16.

Newman, G. E., J. De Freitas, and J. Knobe. 2015. "Beliefs About the True Self Explain Asymmetries Based on Moral Judgment." *Cognitive Science* 39(1): 96-125.
Nichols, S., and M. Bruno. 2010. "Intuitions About Personal Identity: An Empirical Study." *Philosophical Psychology* 23(3): 293-312.
Nichols, S., N. Strohminger, A. Rai, and J. Garfield. 2017. "Death and the Self." MS submitted for publication.
Nozick, R. 1981. *Philosophical Explanations.* Cambridge, MA: Harvard University Press.
Parfit, D. 1971. "Personal Identity." *Philosophical Review* 80(1): 3-27.
Parfit, D. 1984. *Reasons and Persons.* Oxford: Oxford University Press.
Paul, L. A. 2014. *Transformative Experience.* Oxford: Oxford University Press.
Paul, L. A. 2015. ":What You Can't Expect When You're Expecting." *Res Philosophica* 92(2): 149-70.
Peetz, J., A. E. Wilson, and E. J. Strahan. 2009. "So Far Away: The Role of Subjective Temporal Distance to Future Goals in Motivation and Behavior." *Social Cognition* 27(4): 475-95.
Reed, A. 2004. "Activating the Self-Importance of Consumer Selves: Exploring Identity Salience Effects on Judgments." *Journal of Consumer Research* 31(2): 286-95.
Reed, A., M. R. Forehand, S. Puntoni, and L. Warlop. 2012. "Identity-Based Consumer Behavior." *International Journal of Research in Marketing* 29(4): 310-21.
Rehder, B. 2003. "A Causal-Model Theory of Conceptual Representation and Categorization." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 29(6): 1141-59.
Rehder, B., and R. Hastie. 2001. "Causal Knowledge and Categories: The Effects of Causal Beliefs on Categorization, Induction, and Similarity." *Journal of Experimental Psychology: General* 130(3): 323-60.
Rhodes, M., and A. C. Brandone. 2014. "Three-Year-Olds' Theories of Mind in Actions and Words." *Frontiers in Psychology* 5: 263.
Rips, L. J. 2011. "Split Identity: Intransitive Judgments of the Identity of Objects." *Cognition* 119(3): 356-73.
Rips, L. J., S. Blok, and G. Newman. 2006. "Tracing the Identity of Objects." *Psychological Review* 113(1): 1-30.
Rips, L. J., and S. J. Hespos. 2015. "Divisions of the Physical World: Concepts of Objects and Substances." *Psychological Bulletin* 141(4): 786-811.
Rosch, E., and C. B. Mervis. 1975. "Family Resemblances: Studies in the Internal Structure of Categories." *Cognitive Psychology* 7(4): 573-605.
Ross, L., D. Greene, and P. House. 1977. "The " 'False Consensus Effect': An Egocentric Bias in Social Perception and Attribution Processes." *Journal of Experimental Social Psychology* 13(3): 279-301.
Sagi, E., and L. J. Rips. 2014. "Identity, Causality, and Pronoun Ambiguity." *Topics in Cognitive Science* 6(4): 663-80.

Sloman, S. A., B. C. Love, and W. K. Ahn. 1998. "Feature Centrality and Conceptual Coherence." *Cognitive Science* 22(2): 189-228.

Smith, E. E., and D. L. Medin. 1981. *Categories and Concepts.* Cambridge, MA: Harvard University Press.

Strohminger, N. 2015. Neurodegeneration and Identity. *Psychological Science* 26(9): 1469-79.

Strohminger, N., and S. Nichols. 2014. "The Essential Moral Self." *Cognition* 131(1): 159-71.

Swann, W. B. 1983. "Self-Verification: Bringing Social Reality into Harmony with the Self." *Social Psychological Perspectives on the Self* 2: 33-66.

Tobia, K. P. 2015. "Personal Identity and the Phineas Gage Effect." *Analysis* 75(3): 396-405.

Urminsky, O. 2017. "The Role of Psychological Connectedness to the Future Self in Decisions Over Time." *Current Directions in Psychological Science* 26(1): 34-9.

Wellman, H. M., and S. A. Gelman. 1992. "Cognitive Development: Foundational Theories of Core Domains." *Annual Review of Psychology* 43(1): 337-75.

Wilson, A. E., and M. Ross. 2001. "From Chump to Champ: People's Appraisals of Their Earlier and Present Selves." *Journal of Personality and Social Psychology* 80(4): 572-84.

Yang, A. X., and O. Urminsky. 2015. "The Foresight Effect: Local Optimism Motivates Consistency and Local Pessimism Motivates Variety." *Journal of Consumer Research* 42(3): 361-77.

4. Models of Transformative Decision-Making[(6)]

Samuel Zimmerman and Tomer Ullman

> Let not future things disturb thee, for thou wilt come to them, if it shall be necessary, having with thee the same reason which now thou usest for present things.
>
> Marcus Aurelius, The Meditations, trans. G. Lang

1. Introduction

New things happen to people all the time, and there is nothing surprising in that. But people also constantly make decisions about new things, and there is something strange about that. From tasting a novel dish to moving to an unknown city, from going on a date to going to war, how can people reasonably choose the unknown? A rational decision-maker makes her choice by considering the costs and benefits of an outcome and weighing it against alternatives, and she can use simple probabilities to capture simple uncertainties. But what should she make of cases where she does not know the costs and benefits? And what should she do when the decision will alter her very core, the beliefs and desires she considers when making a decision?

Philosophical investigations of how people change their self in light of new experiences are not themselves new (Locke 1700). Some of the oldest philosophy centers on this theme, such as Heraclitus' dictum: "No person ever steps in the same river twice, for it's not the same river and they're not the same man" (Robinson 1987). More recent philosophical work has explored both the difficulty of rationally reasoning about the self (Jackson 1986; Parfit 1984), and the fundamental difficulty of applying standard decision-making frameworks to big, self-altering decisions (Paul 2014; Ullmann-Margalit 2006).

In this chapter, we consider the psychological architecture that supports decisions about novel and transformative experiences, in light of advances in the computational

(6) Samuel Zimmerman and Tomer Ullman, Models of Transformative Decision-Making In: Becoming Someone New: Essays on Transformative Experience, Choice, and Change. Edited by: Enoch Lambert and John Schwenkler, Oxford University Press (2020). © Samuel Zimmerman and Tomer Ullman.
DOI: 10.1093/oso/9780198823735.003.0005

modeling of thought. The first half shows how people can make informed decisions about trying unfamiliar things, by using rational inductive inference to evaluate novel experiences. This inference integrates previous experience, current preferences, and an understanding of how the world is organized, using a non-parametric hierarchical Bayesian model (Griffiths et al. 2008; Tenenbaum et al. 2011) to capture the structural uncertainty in such decisions. The model also accounts for higher-level decisions, by allowing for higher-order preference: people can decide to try something new because they like the novelty of it, regardless of the particular immediate sensation that experience brings.

The second half of the chapter tackles decisions about transformative experiences, meaning choices that can affect people's self, including people's decision-making faculty and the preferences it relies on. Reasoning about such choices requires that people have a theory-of-self, similar to the theory-of-mind they have of others. We present a formal framework for adjudicating between selves based on such a theory- of-self, and explore several alternative models within this general framework. We use both formal modeling techniques and empirical work to compare the different models within these framework, and end by considering what is still left out of a descriptive, computational account of making big decisions.

While we regard the following work as an important contribution to the philosophical and psychological issues around transformative experience, we consider the empirical results as initial explorations, setting up the edifice for future work into attitudes towards real-world transformative experience.

2. Grape Decisions: Decision Theory and Novel Experience

Imagine a philosopher and her friend the layman (that is, a non-philosopher) walking through a street market. The layman comes across a yellow grape-like object he's never seen before, sitting on a counter next to red grapes, blue grapes, and green grapes. The layman picks up the yellow item, thinks for a moment, and pops it in his mouth.

Decisions like these are made every day with barely a second thought. Despite the ordinariness of it (or because of it), the philosopher decides to challenge her friend. How can the layman possibly have decided to try the new fruit, she asks, and claims this decision was made irrationally.

The philosopher argues as follows (the following is based largely on Paul 2014; 2015a). In order for the decision to try a new experience to be rational, it must follow the rules of decision theory. In standard normative models for making decisions under uncertainty, a decision is made by considering all possible outcomes in terms of the likelihood they'll occur, and the possible benefits and harms to the decisionmaker in case they do occur (e.g. see Weirich 2004). The benefit and harm of an outcome can

be captured by a utility function, and a rational decision is that which maximizes the agent's expected utility.

An agent's utility function may be informed by testimony (other people offering explanations, reasons, and self-reports about an experience), observation of others, or an agent's own personal experience. But, argues the philosopher, in certain situations neither testimony nor observation are enough to inform one's utility function (Lewis 1990; Paul 2014). And assuming neither observation nor testimony are adequate with respect to describing the experience of eating a yellow grape-looking thing, the agent apparently cannot assign a utility to the possible outcomes of the decision.

In short, making a rational decision where the utility depends on perceptual experience requires knowing what the resulting experience will be like. If the layman did not know in advance what the yellow item will taste like, he cannot rationally choose to try or decline tasting it. The layman counters by pointing out that while he never tried this yellow grape-looking thing before, he has eaten green grapes, and purple grapes, and red grapes, and liked each of them. He reasonably assumed this new item was a sort of grape, and since he likes grapes, he thought he'd like this new one.

This yellow grape example (in a slightly different form) was previously discussed in the philosophical literature as an example of the difficulty novel experiences present for a rational decision-maker (Paul 2014). We consider it here in order to formalize the intuitions of the layman, and show that his decision is rational, as an example of the way previous experiences combined with a structured understanding of the world can be used to evaluate new items. At the same time, this formalization shows the machinery that underlies commonsense reasoning is not trivial. The layman may be justified in his decision, but the philosopher was justified in calling attention to it.

To spell out the layman's commonsense intuition more before formalizing it, it is an intuition based on an understanding of how the world is organized, and an understanding of the decision-maker's own desires and preferences. That is, the layman abstracted away from particular instances (red grapes, green grapes, etc.) to a more general category (grapes). He then inferred that the new item must be a new sub-item of this general category, based on its observable properties (the yellow grapes look like grapes, so they probably are grapes). The layman further assumed that unseen properties of a new instance are similar to previous instances (a new grape will probably taste similar to previous grapes), and based on his previous preferences was able to infer his likely preference for the new item (since he liked the taste of previously tried grapes, he'll probably like the taste of this new grape). The layman's intuitive understanding is characterized in Figure 4.1.

The following section grounds the intuitive inference about simple novel experiences using a Bayesian hierarchical model. In order to account for the potential introduction of novel items, the model is non-parametric at every level.

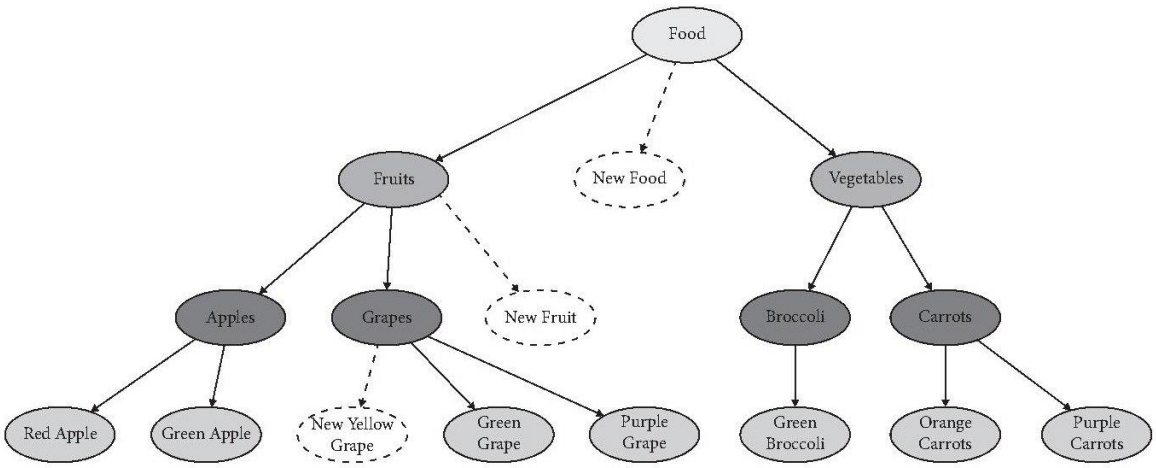


Figure 4.1 A simple intuitive model of the relationship between food items. Each instance of a specific food is at the lowest levels. While color varies within a given food, the shape and taste are less likely to do so. For instance, green and purple grapes share shape and taste, while green grapes and green broccoli share color but do not taste the same. Fruits on the whole have a similar taste, that is distinct from vegetables. A non-parametric hierarchical Bayesian model is able to hypothesize new categories and sub-categories at different levels of the tree.

2.1 Eating From the Tree of Knowledge: Structured Knowledge and Decision Theory

An agent's decision to try something new (such as a yellow grape-like thing) is partly an induction problem. The agent has experienced a number of objects before, and formed an associated preference over them that depends on their properties. The agent now needs to infer both whether a new object is a member of a previous category or a new category, and how the properties of the new object relate to previous objects. Our model is phrased in terms of food, taste, and shape in order to give a concrete example along the lines discussed in Paul (2014), but it applies more broadly to the decision to try or decline novel experiences with properties that depend on their category.

Our model posits a rich set of latent structure to perception that enables *unconscious inference*. Agents structure or process the "blooming, buzzing confusion" of their sensory data to make inferences about the hidden structure of the world (e.g. that there exists such a thing as a grape). Within such structures, agents are able to reason with much less information, and much more quickly, than if dealing with unstructured sensory inputs.

As desiderata, a model that can accommodate a decision about a new item in a potentially new category should:

1. contain nested levels of domain specific information (similar to the structuring of grapes as fruits and fruits as foods).

2. have lower levels inherit their properties from higher, more abstract levels. Thus, more similar categories should have similar prototypes, and also have similar variability in the ways they differ (apples and pears are more informative in making the decision about grapes, compared to broccoli and jalapenos).

3. enable rich covariance relationships between features. For example, some features should be correlated with one another (such as the fruit's shape and its taste), while others should not (such as the fruit's color and its taste).

4. accommodate for novel experiences by enabling a potentially infinite number of items at each level of the structure (this enables a simultaneous inference over whether a new item is a new grape, a new fruit, or even a new type of food).

Note that this list is incomplete, and a model that meets it is a simplification and not meant to capture the full range of decision-making and evaluation when assessing a new item such as a new food. In particular, the list does not mention compositionality of traits.[1] One model that does meet all the above criteria is a hierarchical Bayesian non-parametric model. We next define such a model in detail.

[1] And see Gershman et al. (2017) for a recent treatment of structured utilities that compose properties in the food domain.

2.2 A Model for Choosing a New Item Based on Past Experience

Object Representation

The model represents a specific food object *Fi* as a tuple containing discrete taste, shape, and color attributes:

$$Fi = \{ti;\ si;\ Cig\ ;$$

where the different attributes are Boolean vectors. The taste attribute *ti* is a Boolean 5-vector representation of each of the five major taste buds: sweet, sour, salty, bitter and umami. For example, a food *Fi* that is only sweet and sour will have $ti = (1; 1; 0; 0;$ o). Also, *si* is a vector representing the food's shape, and *ci* is a vector representing color.

Utility Function

According to the model, agents have a utility function *U* that assigns a scalar value to a food object *F*. This utility function is informed by previous experience, and encodes the agent's expected *hedonic* pleasure or pain.[2] We assume for simplicity an agent's basic preference for a food item depends primarily on the taste of the food. An agent's utility function *U* is thus a mapping from the 5-vector *tf* to a number, such that for two food items A and B, if *U(tf (A))>U(tf (B)* then the agent prefers A to B. The utility of objects not previously experienced can be reasoned about probabilistically, by inferring the likely t_{new} of the unknown object F_{new} .

In the running example, the layman derived utility from an item being similar to the grapes he's had before. Thus, we define a simple utility function:

$$Ufood\ (Fi) = \hat{}\ ^{U}taste(^{t}ij)$$

Where U_{taste} is the utility derived from the specific taste in the taste-vector. While we restrict ourselves in this section to utilities that depend on taste and sum similarly over the taste components, other utilities are possible, including non-linear combinations of taste and more abstract utilities such as a preference for or dislike of new and unknown items. Our goal here is not to accurately model the ways in which subutilities combine for taste. We consider these more general utilities in the discussion.

[2] Although this may differ from the actual moment-utility, and see Bentham (1996); Kahneman etal. (2003); Kahneman etal. (1997); Elster and Loewenstein (1992).

Generative Structure of Objects and Properties

The specific food object *F = {tf; sf; Cf g* is an instance at the end result of a generative category hierarchy, going from "food" to "category" (e.g. fruit) to "sub-category" (e.g. grapes) to "instance" (e.g. Concord grapes).

The topmost level, "food," contains hyper-parameters used for drawing new "category" objects. New categories are sampled in the following way, using the "food" hyper-parameters:

tij - Betai^food) Si - Dirichlet(afood) Ci — Dirichlet (afood)

Where *tij* refers to a specific taste component *j* within a taste vector related to food *Fi*. The "sub-category" objects are sampled in a similar manner to the "category" objects, though instead of β distributions over the set of possible tastes, a *Binomial* function represents variability in each taste vector within different instances of the same food "sub-category." This approximately captures the extent to which foods from each category are similar (how much we should expect grapes as a whole to be similar or different in taste). In our model, we specify all of these parameters within a reasonable range to describe the structure of the world before trying a new item, but they can be learned.[3]

The particular taste, shape, and color attributes of a new food are sampled from its sub-category via the multinomial and binomial distributions concentrated with the parameters of the "sub-category" before it. These *6* variables indicate the feature distributions across each food instance (e.g. yellow grapes and green broccoli).

We instantiated the above model in a *probabilistic program* through the CHURCH probabilistic programming language.[4] This generates a set of food instances with shape, taste, and color attributes. This model assumes that the agent is able to solve the inference about where to place the novel food object in the tree. In the technical appendix[5] we provide a method for how, given access to observable properties (i.e. the food's shape and color), an agent is able to infer where to place the new object within the hierarchy, and from there infer its hidden properties (i.e. the food's taste).

Everyday Predictions

The model above generated fruitful predictions about the novel experience of trying the yellow grape. It was conditioned upon the agent having eaten green grapes and green broccoli before, and upon the object under consideration being yellow and grape-shaped. Additionally, we conditioned the sweetness, umaminess, bitterness, and

[3] See Kemp et al. (2007) for an account of a similar model that learns these parameters. This is captured via learning a posterior distribution p(a, 0|n).

[4] For an in-depth treatment on the topic and the philosophy behind it, see Goodman and Tenenbaum (n.d.).

[5] <www.samuelzimmerman.com/transformativeAppendix>

sourness of the green grapes and the green broccoli as not equal (independent of what the taste of each was).[6]

Taste and utility of new item: Given that an agent can place the new food object in the hierarchy (see the technical appendix), the agent is able to form a hypothesis about the taste of the new food, namely that it will probably taste like a green grape, and will deliver a similar utility. From this, the utility of eating the new grape can also be calculated (see Figure 4.2).

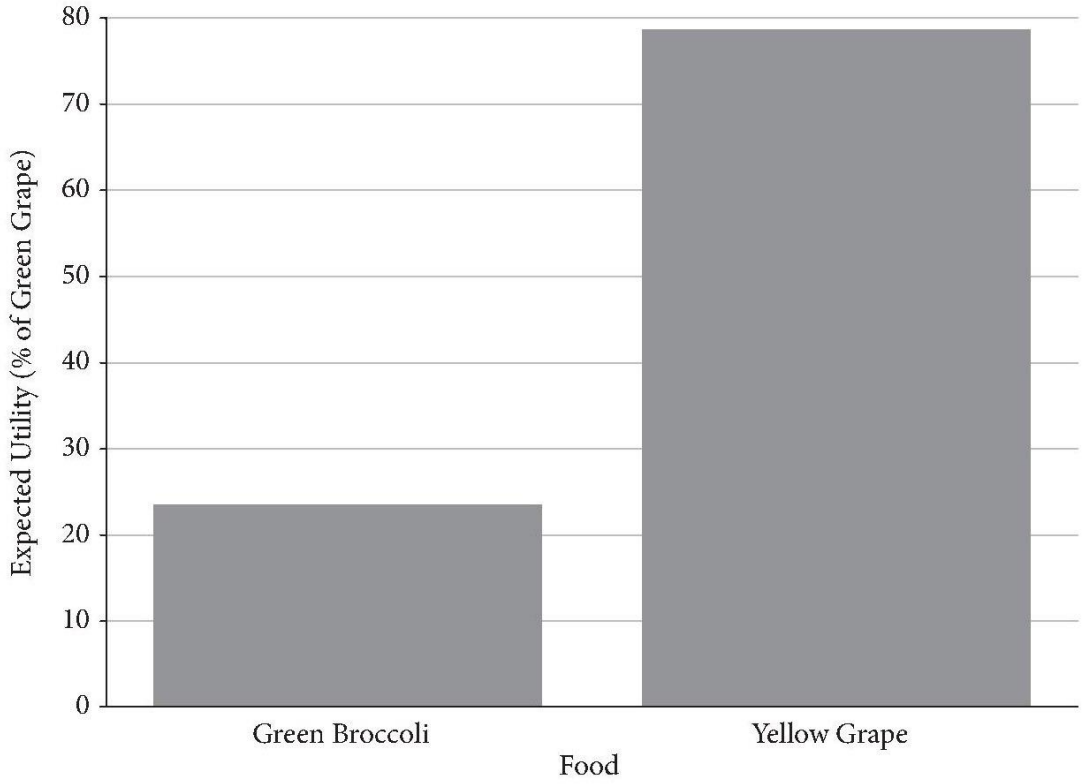


Figure 4.2 The expected utility of the new food object (yellow grape) and a known unfavorable object (green broccoli), relative to the utility of the favorable object (green grape).

In short, even with very little information the above model is able to infer the relation of new objects to previously encoded objects, and then produce an informed hypothesis about the likely perceptual qualities of the novel experience.

[6] For the "category" inference, we had (.11 0.5) for the a parameters on the generative side corresponding to variability in shape, color, and taste.

2.3 Grape Decisions: Discussion

Our model shows how an agent can make a justified decision about a novel experience, using previous perceptual evidence to form a hypothesis about the utilities of experiences they have not had. While the decision is justified, the computations involved are not trivial, and rely on the ability to structure the world, and to account for potentially new objects through non-parametric reasoning. Such everyday, intuitive decisions made through "common sense" or intuition, when prodded by philosophical questioning, turn out to have a rich structure to them. This is similar to the observation that facial recognition is fast and intuitive, but being able to perform it does not mean people have explicit access to the underlying computations that are carried out by the visual cortex when we recognize our friend coming towards us (Kahneman 2011).

Our model considered simple utilities, tied to the particular properties of an object instance (the taste of a food item). But the model can also consider more abstract utilities, relying on the non-parametric hierarchical structure of the agent's knowledge. For example, the agent may derive negative or positive utility from "opening up" novel categories at each level of the model:

$$U(\mathrm{F}) = \mathit{Utaste}\ (F) + \mathit{Unew}\ (\mathrm{F})\ ;$$

where *unew^F)* is positive or negative depending on whether F is a new food instance. That is, regardless of particular taste experience, the very generation of a new experience category might in itself be associated with a positive or negative utility.

Consider the case of an agent deciding whether to try a novel food called *durian* (Paul 2014). Durian is a large and odorous melon native to Southeast Asia. By all accounts, eating a durian is a highly unique experience.[7] As testimony is unlikely to help describe the experience, except that it is unique, how might previous experience inform an agent's reasoning about whether to try such a food?

The agent may derive pleasure (or pain) from creating a whole new category of fruit, or even a whole new category of food. Such pleasure from exploration (or aversion to it) may be related to the "Openness" factor in human personality, as measured in the popular and empirically vetted OCEAN framework (McCrae and John 1992). Our model, coupled with the non-parametric extensions in the appendix, extends this measure by representing a hierarchical notion of novelty itself. One can also imagine valuing novel experiences at higher levels of the hierarchy more, as they provide a more foundational change to one's understanding of the world. So, while a new object may have certain unknown ground features, and may present a problem for a decision-making account that relies on knowing these features, the very fact that the object is

[7] Paul (2014) writes: "One important chef says, 'The only way to describe its taste is 'indescribable.' ... [other] descriptions [include]: 'Eating vanilla ice cream by a sewer' or 'French-kissing a dead rat'" (p. 35).

new presents a higher-order, observable feature which the agent may take into account in its decision-making. We mean this as a normative point about how higher-order decisions can be made, and how they can be captured specifically as the opening of a new node in a hierarchical representation, and do not mean to suggest that descriptively all people will in fact weight novelty for good or ill in their decision-making, nor that novelty is the only higher-order observable feature of new objects.

For all its uniqueness and novelty, the experience of eating a durian for the first time likely leaves the decision-maker essentially unchanged. This is an epistemic change, in that it opens up new possibilities of pondering experiences, but it is not a self-change to the agent's model of itself. A less common, far more gnarly, and much more interesting case is that of transformative experiences that change the agent itself in a more fundamental way. These are experiences where the desires, intentions, qualities, and beliefs that led to the decision are themselves altered. This is the case we turn to next.

3. Who Decides on the Decider? Decision Theory and the Intuitive Theory of Self

How do we understand other minds? A leading view in cognitive psychology is that people construct an intuitive theory-of-mind when reasoning about other people (Dennett 1989; Gopnik 1993; Happe 2003). According to this intuitive theory, other agents have beliefs, desires, intentions, and various other mental states that lead them to take certain actions. While these desires and beliefs cannot be directly observed, we can infer them from how people act, and use them to explain past actions and predict future actions. This inference process has been formalized in recent years as "Bayesian theory-of-mind," and used to capture adult and children's reasoning about the social and non-social goals, beliefs, and relations of others (Baker et al. 2017; Baker et al. 2009; Baker and Tenenbaum 2014; Hamlin 2013; Jara-Ettinger et al. 2016; Jern and Kemp 2015).

How do we understand our own mind? A constructed notion of an essential "true self" as the binder of intentions, desires, and in particular moral traits has been explored in cognitive psychology (e.g. Newman et al. 2014; Strohminger and Nichols 2014; Strohminger etal. 2017). But what would a formalization of an intuitive "theory-of-self" look like? It may be quite similar to how theory-of-mind is constructed for other people, informed by additional (though not full) access to memories and internal states (Baker and Tenenbaum 2014; Saxe 2009). Note that on such a view it is not necessary that the theory-of-self truly aligns with the self, any more than the theory of intuitive physics truly aligns with real physics, or theory of mind aligns with other people's *actual* decision-making, which is the subject of decision theory and behavioral economics. The deliberations and explanations that a person gives to themselves when

contemplating a self-changing decision may be quite different from the actual reasons that drive them to make a decision.

How do we understand a transformative experience? Ullmann-Margalit (2006) defined transformative decisions as those that abruptly alter the core desires and beliefs of the decision-maker.[8] A transformative decision is one that takes us from our current model of our self (as an agent with particular desires, intentions, and beliefs) to a different agent (with new desires, intentions, or beliefs). Transformations that happen with discontinuities of space or time frequently call into question a notion of self (such as the teleportation experiment—see Parfit 1984). To this we can add discontinuities in a more abstract space of beliefs and desires.

Formalizing the intuitive theory-of-self along the same lines as Bayesian theory-of-mind casts the deliberation about transformative decisions as an agent-based decision-making problem, and gives the ability to specify quantitative differences between different selves. That is, the deliberation can be seen as a problem of expected utility maximization, where the overall utility is defined over agents (current and transformed) that themselves have different utilities and beliefs. But such a formalization leaves open many questions about how the decision should be made: What overall utility function should be used to arbitrate between different agentspecific utility functions? Should the decision-maker construct an external metaagent to arbitrate between their current self and their potential future selves? Should an agent's current utilities and beliefs matter more than the potential future self, and in what way? Can one even make a rational decision about a future self, in the same way that one makes a rational decision to eat a yellow grape?

These questions correspond to questions raised by philosophers examining transformative experiences (e.g. Paul 2014). Pettigrew (2015) in particular recently showed that a transformative decision can be "rational," in the sense that it can be made within an appropriate decision-making framework. In Pettigrew's framework, local utilities are assigned to successive chunks of time, and weights are assigned to each utility. The question then becomes that of which weighting one wishes to place on different utilities (perhaps you care more about the utilities of your current self, or perhaps you discount utilities after a transformation). Pettigrew himself points out that this leaves open the question of how to set the weights, but the problem is at least rationalized. However, as Paul points out, this framework only treats the decisionmaking problem from the point of view of the current self Sc (Paul 2015b). From the point of the transformed self S_t, the utilities will be assigned a different weight, and there is perhaps no rec-

[8] As Ullmann-Margalit (2006) puts it: "[L]et us think of cases of opting as cases in which the choice one makes is likely to change one's beliefs and desires (or 'utilities'); that is, to change one's cognitive and evaluative systems. Inasmuch as our beliefs and desires shape the core of what we are as rational decisionmakers, we may say that one emerges from an opting situation a different person. To be sure, there is a sense in which every choice changes us somewhat. The accumulation of these incremental changes makes us change, sometimes even transform, as life goes on and as we grow older. But what I am here calling attention to are the instances in which there is a point of sharp discontinuity" (pp. 158-9).

onciling the two. Imagine for example an artist considering becoming a lawyer. The artist imagines going to lawyer parties and assigns the derived utility a low weight, for various reasons. However, if she were truly to become a lawyer, from the perspective of the transformed self, these same parties might be assigned a high weight. One cannot then say that the artist is choosing rationally to not become a lawyer, from the perspective of the lawyer. But how is the artist to choose? Pettigrew considers placing second-order utilities over utilities, but rejects this as leading to an infinite regress of higher-order utilities. However, this worry should be embraced as part of the problem of transformative decision-making. Furthermore, it may be that in practice people only consider a small or single transcendence in utility functions.

In the following sections, we describe a common framework for agent-based decisions, and then formalize a simplified framework for a theory-of-self, along the lines of a model for theory-of-mind for other agents. We show the different possible models an agent may have for arbitrating between different selves. We evaluate these different models by considering their predicted output, and comparing it to the empirical behavior of participants in four experiments.

3.1 Agent-Based Decision Theory and Simple Decisions

We assume that people explain the behavior of others by thinking of them as agents acting rationally to achieve their goals, subject to their beliefs about the likely state of the world (Baker et al. 2011; Jara-Ettinger et al. 2015). Again, this view is agnostic as to the mechanism people *actually* use to make decisions. Rather, this is a formalization of the intuitive theory people use to explain and understand, to others and themselves, the causes of their decisions.

Following standard planning algorithms, we assume a world that is in a particular circumstance c out of a set of possible circumstances C (see e.g. Russell and Norvig 1995).[9] At each point in time an agent can take an action a out of a set of actions A, the result of which is a new world circumstance c. A transition function T defines the probability of moving from one circumstance to another circumstance, given an action:

$$T(c,\ C;\ a) := \mathrm{P}(\mathrm{Ct{+}1} = cjCt = c,\ At = \mathrm{a}).$$

Agents derive value from achieving their goals in a given world circumstance $C = c$, or (possibly ordered) set of world states ci 2 C, which we can represent through a utility function $U(ci)$.

An agent does not necessarily have direct access to the true circumstance of the world, or to the true transition function. Rather, the agent can make observations of the world o 2 O, and use the observations to inform and adjust their belief about the current circumstance. An agent's belief about the current circumstance, and the likely

[9] The common terminology uses "state" instead of "circumstance," but the variable S will soon be used to indicate different selves.

transition function given those observations, are represented via functions B(C|O) and *B(T*|O).

Agents have a planning procedure, *PLAN*, which takes in the agent's utilities, beliefs, and possible actions, and returns a probability distribution over the set of actions the agent can take:

$$PLAN\ (\ U\ ;\ B\ ;\ A\ ;\ C\ ;\ O\hat{}\hat{}\ P\ (A = a\hat{}\ ;\ B\ ;\ O = O)\ ;$$

Where *P(A* = a| U*; B)* is the probability of taking a particular action *a,* given utilities *U*, beliefs *B*, and observations *o.* A rational planning procedure returns a distribution over actions such that the agent acting in accordance with this distribution will maximize its expected utility. There are many different ways to implement PLAN, such as Markov Decision Processes, Partially Observable Markov Decision Processes (Kaelbling et al. 1998), and Planning-as-Inference (Todorov 2004).

By assuming that agents implement a planning procedure (see Figure 4.3) that produces actions in order to maximize utility under constraints, an outside observer can use inverse Bayesian reasoning to infer the latent variables (the likely beliefs and utilities of an agent) given the observed variables (the agent's actions and the world):

$$P(U,B \mid A = a,\ C = \mathrm{c})\ /\ P(A = a \mid U,\ B,\ C = c)P(U,\ B),$$

where *P(U, B)* is the prior belief the observer places on the beliefs and utilities of an agent, and $P(A = a \mid U,\ B,\ C = \mathrm{c})$ is given by a planning procedure as explained above. This is the basic notion of Bayesian Theory of Mind (Baker et al. 2017). In what follows, we focus mainly on the representation of selves as decision-making agents, and less on the inverse-planning aspect of Bayesian Theory of Mind.

3.2 Transformative Choices and an Intuitive Theory-of-Self

We can use the agent-based decision-making framework described above as a starting point for an intuitive theory-of-self. The current self (S_c) is a decisionmaking agent with structured beliefs, goals, intentions, and other mental qualities that drive its actions. The potential transformed self (S_t) is also an agent, with possibly different beliefs, goals, intentions, and mental qualities.

The problem of transformative experience can be recast as a decision-making problem. Given the choice of staying as the current agent *Sc* or changing to a new agent S_t, what should a rational person choose? Note that in this formalization the different agents have different utilities, but this can be recast as agents placing different weights on the same utilities in Pettigrew's (2015) framework.

On a naive "View from Nowhen" account, the agent can simply apply standard expected utility theory to this dilemma.[10] That is, the agent considers the likely future

[10] Cf. *The View from Nowhere* (Nagel 1986). We thank John Schwenkler for suggesting this term.

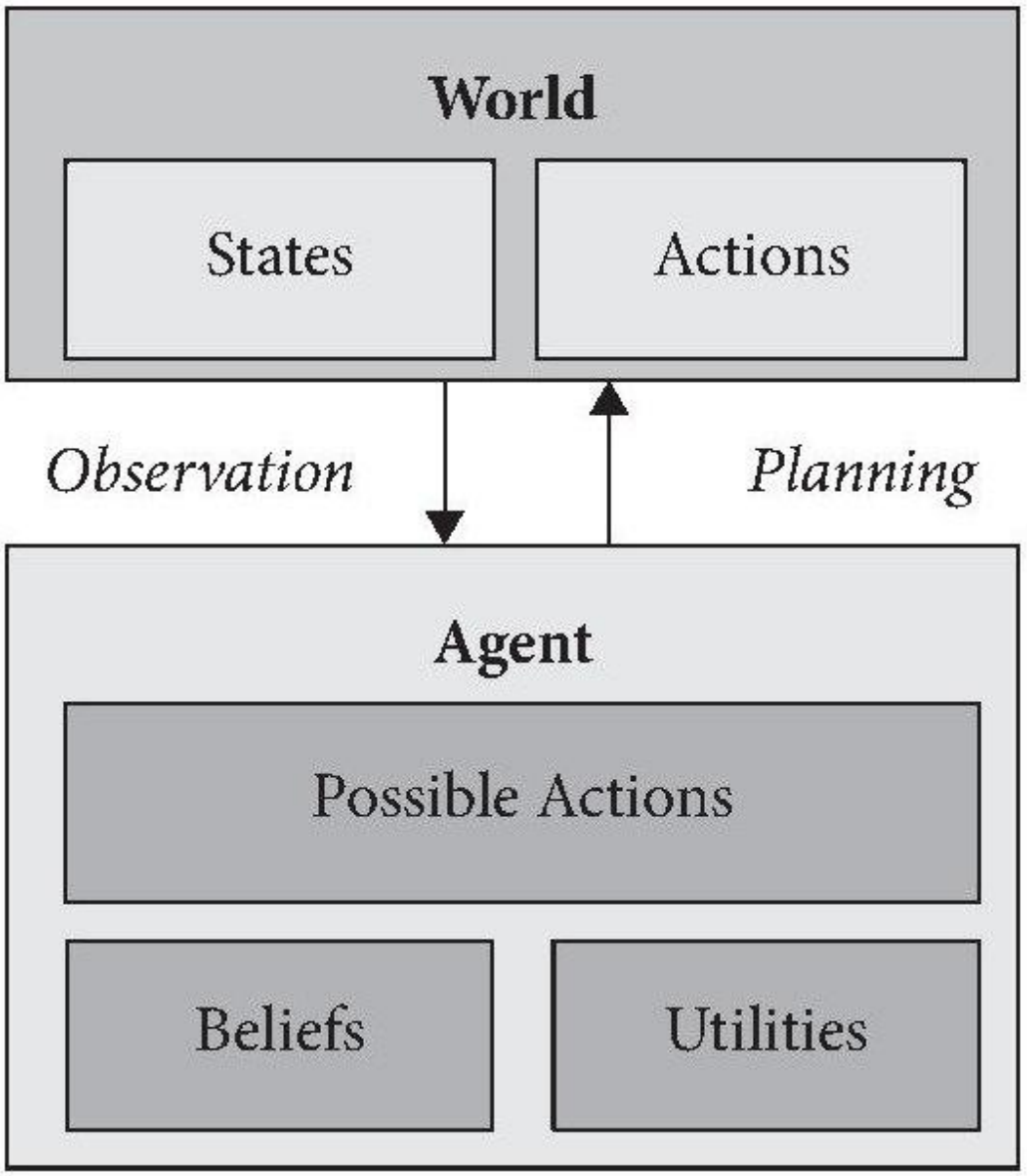


Figure 4.3 This is a simple model for an agent's intuitive planning procedure. Agents' beliefs, utilities, and possible actions receive inputs from observations they make about the world. This informs the actions an agent takes in the world.

circumstances it will encounter as S_c and S_t (different selves may encounter future circumstances with different probabilities), and compares the expected utility from these circumstances under the utilities of the current self *Sc*, and under the utilities of the transformed *St* . The agent then chooses the self who is most likely to be maximize their (expected) utility. In a sense, the agent would be constructing a meta-agent deciding on the meta-action of what agent to be. The meta-agent would map from the set of possible agents A to a particular agent, as shown in Figure 4.4. If we restrict ourselves for simplicity to two agents (current and transformed), we have:

$$MetaAgent(S_c,\ S_t) = \max E(U_S\ |S = Si) = \max\ Si^{\ i}\ Si$$

J2 $P(c\,|\,USi,\ BSi)$-**U**s.$(c),\ c$

where $U_Si\ (c)$ is the utility agent *Si* gets from the state of the world c, and $P(c\,|\,U_Si,\ B_Si)$ is the probability of the world being in state c, given that an agent is *Si.* That is, $P(c\,|\,U_Si,\ B_Si)$ takes into account the particular beliefs and goals of *Si,* and thus its likely actions and resulting state of the world. One could include other mental properties and qualities beyond beliefs and utilities, but these are the common ones explored in theory of mind research.

There are at least two major difficulties with this formulation. The first hurdle is the one explored in Section 2—the difficulty or impossibility of imagining certain future states. As we saw, it is partly possible to get around this difficulty by considering that circumstances and experiences in the world are structured in nature, relying on past experiences and considering meta-utilities on things like novelty.

Such a formulation does not fully solve the epistemic concern, but this difficulty is at any rate not our focus here.

The second hurdle, which is our focus, is that even if it is possible to imagine future circumstances and utilities under a transformed self, it is unclear whose utilities should count. Under a simplistic view-from-nowhen account, such as that expressed in the equation above (MetaAgent), it may be perfectly reasonable to accept the following suicide-for-happiness bargain: your current person is pulverized to a pulp, and the matter is used to reconstruct a new person, who is guaranteed to be happier than you in every way in every circumstance. Most people would reject this idea not because of a failure to simulate the new person, but because the new person is not them. In this suicide-for-happiness bargain, the destruction of the self is physical. But it may be that deep changes to beliefs, utilities, and other aspects of the self also count as a potential self-destruction. The utility of the stranger, the transformed self, does not matter to the current self—and what's more, the stranger is not the one currently making the decision. But an asymmetry is not apparent in the equation.

What other view could people adopt, other than a view-from-nowhen? At least for preferences, one could construct a non-symmetric meta-agent for whom the preferences of the current model-of-self are taken more into account. Informally, in this model, as

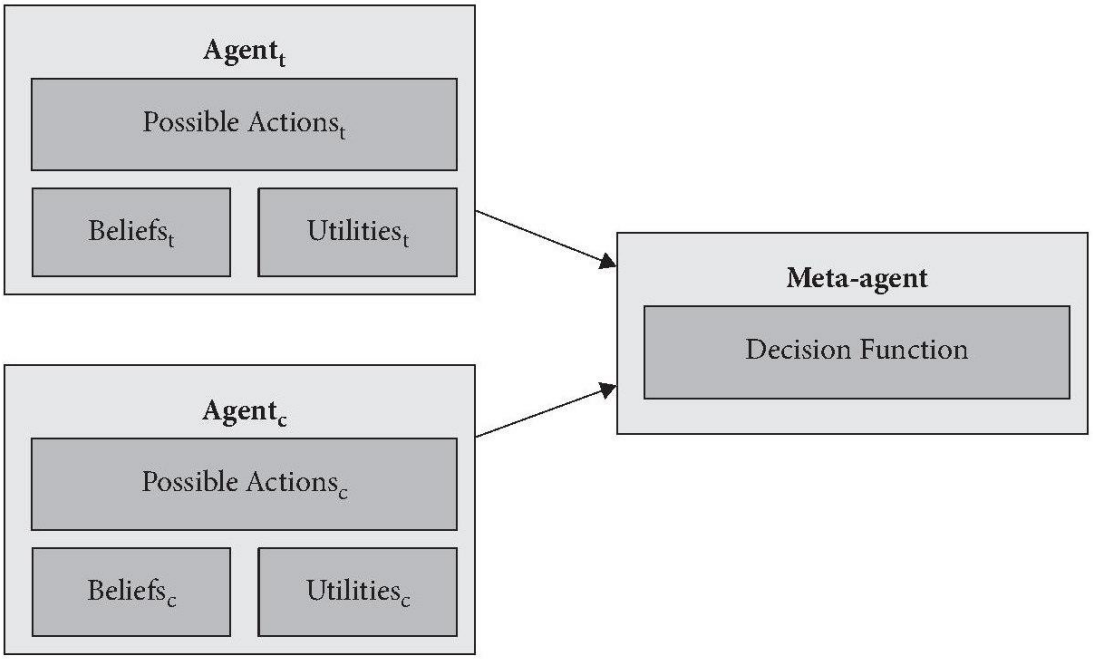


Figure 4.4 We can think of transformational decision as decisions a "meta-agent" makes reasoning about our current self and future selves.

the agent calculates the expected utility of its new self, it asks: "What would my current self think of this circumstance?" even though their current self would no longer experience this circumstance. For concreteness, imagine an aspiring artist contemplating becoming a lawyer, as her demanding father wants. The artist knows that once she becomes a lawyer, she will take part in a standard lawyer's life and will probably even enjoy it. The artist pictures herself in a year's time going to a dinner party with other lawyers, and finds herself bored by the very idea. Alternatively, she imagines going to an exhibit opening at a small gallery, and is quite happy with the thought. The asymmetry is that she is thinking of the dinner party from the point of view of her current preferences—the artist trapped in the lawyer's body. For the exhibition, she is not considering a trapped lawyer-self and their view on the circumstance.

The decision we consider is again made by a meta-agent, but a meta-agent that calculates $\mathrm{EU_i}$ with an eye towards the current self. More formally:

$$\textit{Biased Meta Agent}(S_c; S_t) = \max_{Si\ c} \mathbf{X}\mathrm{P}(\mathrm{c}|\mathrm{U_{Si}};B_{Si})[(1 - a)\text{-}U_{Si}(\mathrm{c}) + a\text{-}U_{Sc}(\mathrm{c})]$$

where US_c are the utilities of the current self. The parameter a controls the degree to which the utilities of the current self are taken into account compared to the utilities of other selves, and may be different for different people. If a is set to 0, we recover the view-from-nowhen in equation. If a is set to 1, the agent parochially considers future events only from the point of view of its current preferences, without taking into account the possibly different preferences of the new self.

Which of these views is more accurate in describing how people act? We first consider the difference between the view-from-nowhen (*MetaAgent*) and the view- towards-self-preferences (*BiasedMetaAgent*) through two studies that ask people to consider a change of their desire.

3.3 Study 1: A Simple Increase of Utility

As a warm-up study, we asked participants to consider a choice between a pleasant and unpleasant experience. Before being asked to choose which experience they preferred to go through, they were given the option of increasing the degree to which they enjoyed the pleasant experience. We expected most people to make use of this option, as predicted by both a view-from-nowhen model, and a view- towards-self-preferences model.

Methods

Eighty participants were recruited on Amazon's Mechanical Turk Service (56 male, 23 female, median age = 30, ranging from 21 to 70 years old; 1 participant failed a catch question and is not included in the analysis).

Participants were asked to imagine that they are facing two doors. If they open the first door, they will be given a mildly painful electric shock that lasts for five minutes and leaves no lasting effects. If they open the second door, they will be given a fluffy puppy to pet for five minutes. Participants were asked to rate the relative pleasantness or unpleasantness of the two experiences, ranging from -100 *(extremely unpleasant)* to 100 *(extremely pleasant).* After rating the two options, participants were asked to choose which door they preferred to open. Participants were then told that prior to opening the doors, they could press a button that would make the pleasant experience more pleasant (+20 points on the pleasantness scale, capped at a maximum of 100). All effects of pressing the button were said to disappear after five minutes. Participants indicated whether they would press such a button prior to opening the doors. They then explained their reasoning, and provided their age and gender.

Results

Participants unsurprisingly rated the anticipated experience of petting the puppy as pleasant (˜+7o), and the experience of the electric shock as unpleasant (—65) in all experiments. Participants indicated they preferred the puppy to the shock in all experiments. All participants went with their indicated preference. Most participants who went with their preference chose to press the button before doing so (71%, 95% CI = 62%-8o%). This result indicates participants are able to reason about their own simple experiential utilities, and are willing to change their preferences. We did not expect nor find any significant gender effects in this or the other experiments. This result is expected both under an unbiased meta-agent, and from a biased meta-agent. In this case, the increase in utility for the transformed self was in line with the preferences of the old self. It is interesting, however, that not all or even a large majority of participants were willing to change their utilities, indicating that participants may place value on having the specific utilities they currently have. Some comments were elucidating in this regard: "I don't want it to be enhanced, it's like taking a drug," "I don't need [the experience] to be any more pleasant," "It's already very pleasant," and so on. We return to this point in Study 4. We next consider a reversal of utilities.

3.4 Study 2: A Simple Reverse of Utility

Methods

We recruited 8o participants on Amazon's Mechanical Turk Service (42 male, 22 female, median age = 30, ranging from 21 to 6o years old; 16 participants failed a catch question). Participants from Study 1 were excluded from participating in Study 2.

Participants were asked to imagine a situation similar to Study 1 (a choice between petting a fluffy puppy and receiving a mildly painful electric shock). As in Study 1,

participants were asked to rate the relative pleasantness or unpleasantness of the two experiences, ranging from -100 *(extremely unpleasant)* to 100 *(extremely pleasant),* and to indicate their choice of door. Participants were then told that prior to opening the doors, they could press a button that would make the unpleasant experience even more pleasant than the pleasant experience (e.g. if "painful shock" was rated as "-20" and petting the puppy was rated as "+30," the shock would now be as pleasant as "+50"). All effects of pressing the button were said to disappear after five minutes. Participants indicated whether they would press a button prior to opening the doors. They then explained their reasoning and provided their age and gender.

Results

In this experiment, 31 of the 64 participants chose to press the button prior to opening the doors (48%, 95% CI = 39%-58%). The z-statistic of the difference between experiments 1 and 2 was 2.73, indicating statistical significance.

Discussion

As opposed to Study 1, only about half of participants are willing to press the utilityaltering button. Such a result is not in line with an unbiased meta-agent, nor is it in line with a view maintaining that experiential utilities are best kept unchanged.

The biased meta-agent (Equation [BiasedMetaAgent]) is in line with these results, with a « 0.1. That is, the utilities of the current self matter, but to a small degree. Note that this analysis does not disentangle the possibility that the difference in decision-rules is happening at a population level. That is, it is possible that about 90% of the population uses an unbiased meta-agent for such meta-preference decisions (a = 0), while the other 10% uses an extremely current-biased meta-agent (a = 1). It is also possible that some people are expressing a meta-preference for not having their preferences changed, regardless of the content of the preferences and regardless of simulating a future self from a particular view-point. This added component is discussed in Study 4.

3.5 Study 3: A Change of Belief

So far we have considered changes of preference, and found evidence in support of people continuing to take their own preference into account even after a transformation of preference occurs. It is as if people are saying, "If I push the button, I will want to choose the shock. And I do not like being shocked." But what of changes in belief? Do the beliefs of the current agent also mix in with the beliefs of the new agent?

In equations (*MetaAgent*) and (*BiasedMetaAgent*), it was necessary to compute P(c| $U_S i$, $B_S i$), the probability the world will be in state *c* given that the agent is *Si.* This can be used to predict that if our future self believes a particular outcome is behind

the left door, and if our future self likes that outcome, then our future self will open the left door. But it is possible that our future self is mistaken about the actual outcome of their action. Perhaps right now, our current self believes the left door is empty, or has something horrible behind it. Even if we imagine a future self with a different state of belief, we consider it a *false* belief (similar to the way we can consider false belief in others (Perner et al. 1987)). Beliefs might hold a different status than utilities for imagining future selves, in that we can imagine ourselves with certain arbitrary utilities (e.g. you like vanilla, but you can imagine yourself liking chocolate, without that preference being a "false" preference), and we can imagine ourselves with certain beliefs, but we cannot hold those beliefs to be the actual state of the world (Shoemaker 1995).

We next consider how people might take into account a different belief of a future self.

Methods

We recruited 80 participants on Amazon's Mechanical Turk Service (42 male, 22 female, median age = 31, ranging from 20 to 68 years old; 16 participants failed a catch question and are not included in the analysis). Participants from Studies 1 and 2 were excluded from participants in Study 3.

Participants were asked to imagine that they are facing two doors. If they open one door, they will be given a mildly painful electric shock that lasts for five minutes and leaves no lasting effects. If they open the other door, they will be given a fluffy puppy to pet for five minutes. Unlike Study 1 and Study 2, participants were told they did not know which door leads to which outcome. Participants were asked to rate the relative pleasantness or unpleasantness of the two experiences, ranging from -100 *(extremely unpleasant)* to 100 *(extremely pleasant).*

After rating the two options, participants were asked to choose which option they preferred to occur. Participants were then informed that prior to opening the doors, they could press a button that would make them absolutely certain the pleasant experience was behind the door on the left. It was emphasized that pressing the button would not reveal the location of the outcomes, but merely change the certainty of the participants. All effects of pressing the button were said to disappear after five minutes. Participants indicated whether they would press such a button prior to opening the doors. They then explained their reasoning, and provided their age and gender.

Results

In this experiment, 24 of the 64 participants chose to press the button prior to opening the doors (32%, 95% CI = 24%-4i%). The z-statistic between experiments 2 and 3 was 1.98, bordering on significance.

Discussion

The majority of participants rejecting the pressing of a belief-altering button cannot be explained by an unbiased meta-agent that holds the beliefs and utilities of all agents equal. On the view-from-nowhen account, the expected utility for an agent that knows the location of a preferable option is strictly greater than one with uncertainty about the location of a preferable option. A simple preference-biased meta-agent that overweighs the preferences of the current agent would also lead to pressing the button, and so also does not account for the result. It is also possible that people prefer not to have their beliefs and preferences manipulated at all, regardless of the content—a point which we consider in the next section.

This result is possibly driven in part due to the fact that while people can conceptualize having different beliefs to themselves, as well as acting on those (presumably false) beliefs, they do not think that their actual beliefs regarding the transition function of the world is wrong.[11] Thus, if they currently believe something bad is behind door A, then they can imagine believing otherwise (that it is good), and they can imagine acting on that false belief (opening the door), but when predicting what would actually happen as a result opening the door, they maintain their current belief (something bad would happen).

This reasoning can be captured by adjusting the previous self-centered meta-agent to take into account the beliefs of the new self St for calculating the likely actions of the new agent, but to use the transition function of the current self Sc for calculating the actual outcome.

More formally, we can break up the term P(cj *USi, BSi*) into:

$$\text{P(c|USi; } BSi) = \text{ˆ}Ps \text{ (cjc'; } action)\text{-}P(actionjUsi, Bsi),$$
$$\text{Action}$$

where *PSi (c | c, action)* is the transition function of the world as agent *Si* believes it to be, and $P(action \mid U_S i; B_S i)$ is given by agent Si's planning algorithm. Instead of using this term, it is possible to use:

$$P(c \mid USi; BSi) = \text{ˆˆ}Psc(c \mid c; action)\text{-}P(action \mid Us,;Bsi); action$$

Notice the move from the transition function PS_i to PS_c . That is, *Si*'s actions are predicated on their (possibly false) beliefs about the world, but the actual result of their actions is dictated by what the current self believes is going to happen.

This is an agent that can imagine a future self being delusional and making bad choices (under the beliefs of the current self), but still imagines the result of those choices as happening according to the current self 's belief. Such a self-centered metaagent would be agnostic about pressing the button, as the expected utility for the agent is the same whether they push the button or not.

[11] Cf. Shoemaker (1995).

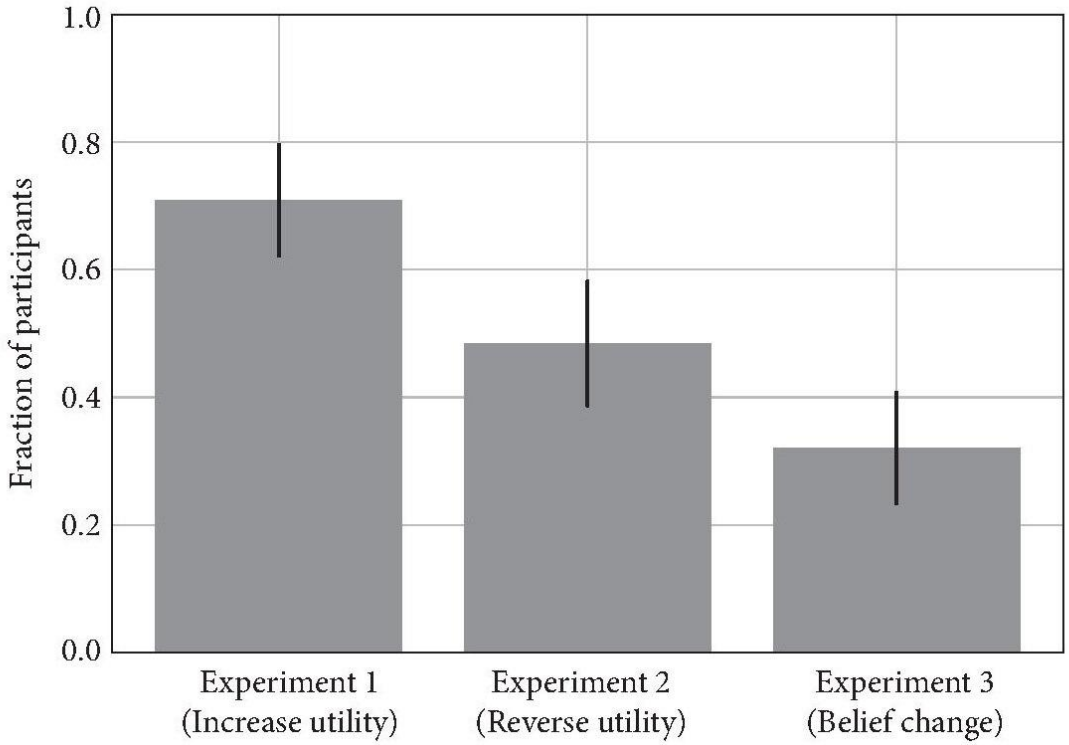


Figure 4.5 Fraction of participants choosing the transformative option in Studies 1-3. Black lines indicate 95 % confidence intervals.

Empirically, however, people were not agnostic between pushing the button and not pushing the button. There are several possible reasons for why this might be the case, including preference for inaction over action (Thaler and Sunstein 2008). One possibility is that people reject changing their beliefs when there is no reason to do so, acting on the "principle of rational belief" according to which there should be a causal link between observations and beliefs (Baker etal. 2011). Again, this would point to the asymmetry between changing utilities (with preferences being in some sense arbitrary) and changing beliefs (people have the beliefs they do for a reason).

3.6 Study 4: A Jump in Self-Space

It appears people value having the beliefs and utilities they currently do, regardless of what the particular values of those are set to. There may be some inherent cost (or benefit) in changing from one self to a different self, not captured by the metaagents considered so far.

To formalize this intuition, we can define a distance metric between selves, as the distance between the belief functions and utility functions of the current agent and the new agent. This distance can then be associated with a utility function (positive or negative). When faced with a decision about whether to change its beliefs and utilities, an agent also thinks about how much it is going to change itself in making the decision.

In general then, we expect:

$$\textit{Distance Meta Agent}(S_c, S_t)$$
$$= \mathbf{X}\mathrm{P}(\mathrm{c}|\mathrm{Usi},\mathrm{Bsi})[(1 - \mathrm{a})\text{-}\mathrm{Usi}(\mathrm{c}) + \mathrm{a}\text{-}\mathrm{Usc}(\mathrm{c})] + \textit{Ud}(d(Si, Sc), c$$

Where $d(Si, Sc)$ is the distance between the potential self Si and the current self S_c. Ud is the utility that the current self Sc places on the distance d. There are many ways for defining a distance between functions, or between distributions. The exact implementation does not matter for the general argument of placing a utility over having a particular utility, a particular set of beliefs, or a trait.

In the next experiment, we consider changing the trait and utility of a person, and the resulting decision problem this poses. We assumed that decisions of changes to the self are partly driven by difference to the self and partly by utility, and so we examine positive changes to a quality while decreasing hedonic utility, and vice versa. We predict people will not choose on the basis of increase in happiness alone, and that their choice will be related to the amount of perceived difference to their self.

We consider changes to intelligence in particular, as in a different experiment (not reported here) we found that changes to intelligence produce large estimations of changes to the self. While "moral qualities" are reported as important to self image, self-selected changes to them may also carry a moral weight which we do not pursue in our current analysis (Strohminger and Nichols 2014).

Methods

We recruited 120 participants on Amazon's Mechanical Turk Service (69 male, 37 female, median age = 31, ranging from 21 to 59 years old; 11 participants were dropped from analysis for misunderstanding a question).

Participants were asked to imagine a device that could change their intelligence, making them more or less intelligent by several degrees. However, the device would also change their happiness, such that becoming more intelligent would make them less happy, and vice versa (the effects ranged from much less intelligent/much more happy to much more intelligent/much less happy, including a "no change" effect). Participants were asked to indicate their own intelligence level on a 10-point scale (5 was population average), and then choose the effect they desired. The exact phrasing of the question was:

> IMAGINE that there was a Device that works as follows: The Device can make you more intelligent, or the device can make you less intelligent. The Device has magnitudes (a little, a medium amount, a lot). If you become more intelligent, you will also become less happy. The more intelligent you become, the less happy you will be. If you become less intelligent, you will also become more happy. The less intelligent you become, the more happy you will be.
>
> QUESTION: How do you use the Device? Do you choose to become ...

Participants were then asked to explain their reasoning ("Why did you choose the way you did?")

On a new form, participants were asked to indicate how different to themselves they would be for different levels of change to their intelligence (regardless of happiness). The intelligence changes were as before, and for each change participants indicated the expected change using a slider ranging from o (completely the same) to 100 (completely different). The exact phrasing was:

> Regardless of happiness, how different do you think you would be from your current self, if you were more or less intelligent? For each level of intelligence change, please select the level of difference to your current self. The answers go from o (exactly the same) to 100 (completely different).

Finally, participants provided their age and gender.

Results

Participants generally rated themselves as more intelligent than average (78%, median intelligence rating of 7). Participants saw changes to intelligence as changes to their self, and a larger increase or decrease in intelligence corresponded with a greater

change to self. As shown in Figure 4.6, participants rated a negative change (decreasing intelligence) as a greater change to the self than a positive change (increasing intelligence).

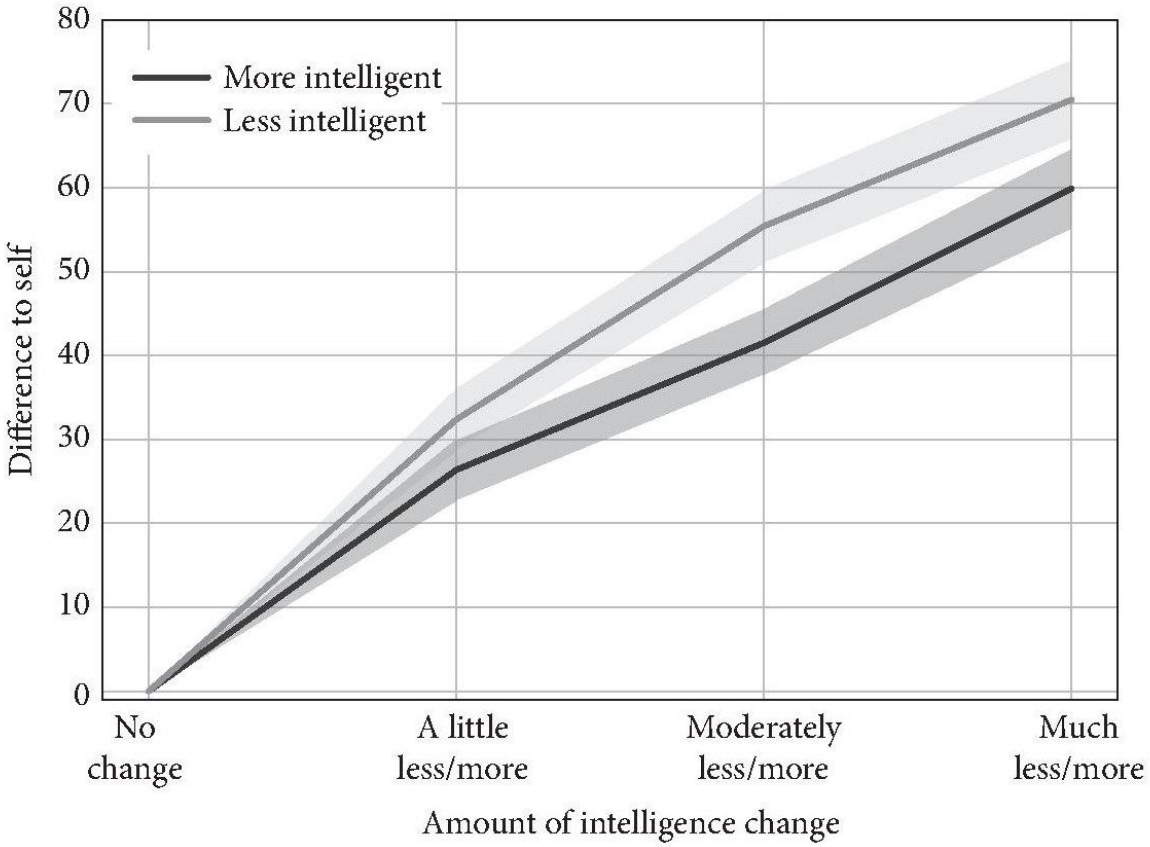


Figure 4.6 Difference to self ratings for different amounts of intelligence change (either increasing or decreasing). Shaded areas show 95% % confidence intervals. Difference increases with amount of change, and becoming less intelligent is seen as a bigger change than becoming more intelligent.

Regarding their decision to change, 50.5% of participants preferred to change nothing about their intelligence and happiness (N=54), 28% preferred to decrease their intelligence and become happier (N=3o), and 21.5% preferred to increase their intelligence and become less happy (N=23). Participants' assessment of their own intelligence was not related to their preference to increase or decrease their intelligence.

As shown in Figure 4.7, participants' estimation of the degree of change they would undergo was related to the direction in which they would change. Considering participants who would prefer to change, those who prefer to become less intelligent (and more happy) see decreasing intelligence as less of a self-change than those who prefer to become more intelligent (and less happy).

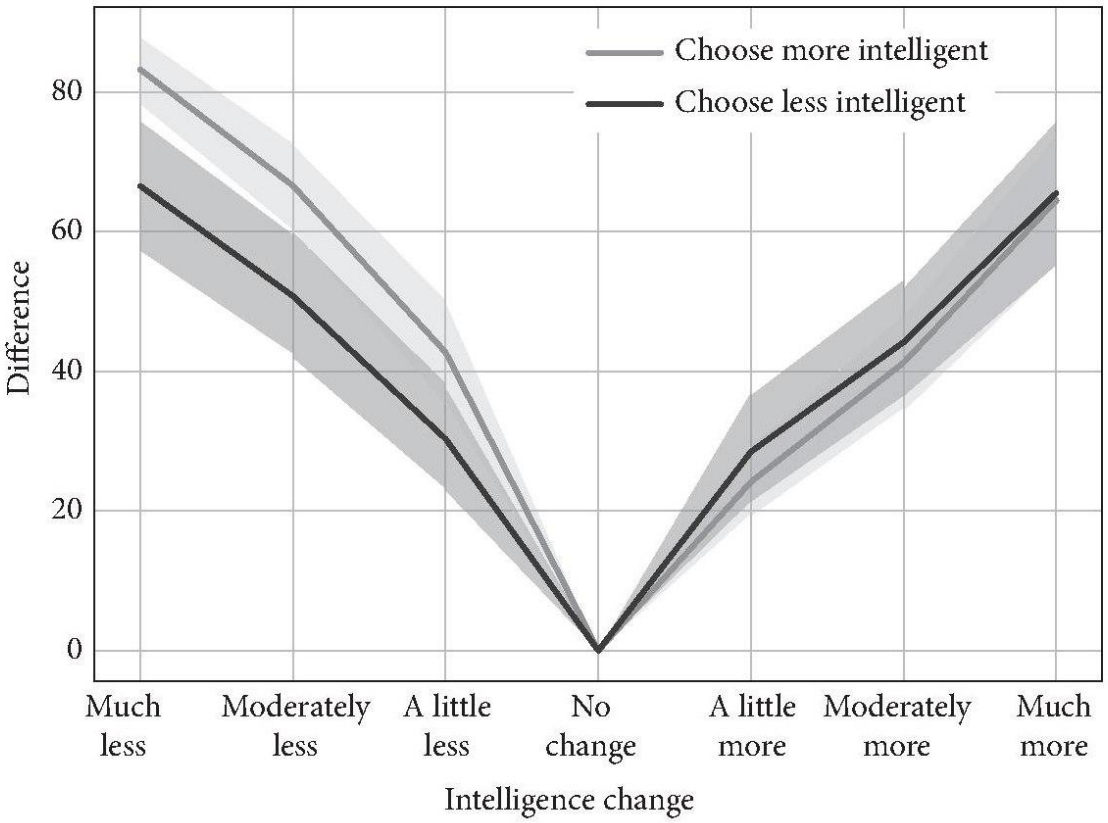


Figure 4.7 Participants' ratings in Study 4 of difference to their self for varying amounts of changes to their intelligence, split by whether a participant chose to become more intelligent or less intelligent. People who choose to become less intelligent see decreases in intelligence as less of a change to their self.

Discussion

On the whole, a majority of participants preferred to remain "as they were," changing neither intelligence nor happiness. Such a change is not predicted by a naïve model of choice, according to which the choice that brings greater happiness should be preferred.[12] We also find that participants in general see a negative change to their self as a greater change. This finding is in line with Strohminger etal. (2017), according to which changes for the better are seen as more in line with one's "true self." However, a majority of participants did not wish to change to be more in line with their true self, when it came as a cost of happiness, and a change to their actual self. Again this is partly in line with Strohminger etal. (2017), who argue that expected changes towards the true self are best seen as a gradual process over time.

We found a relation between the decision to change one's intelligence and happiness in a particular direction, and the perceived difference to the self due to intelligence change. More specifically, participants who would change their intelligence for the worse (to become happier) also see negative intelligence change as less disruptive to their self, compared to those who would change their intelligence for the better. This suggests that beyond simple gained utility, the difference to one's self may also play a role.

Participants' comments were further illuminating. Those who chose not to change expressed a satisfaction with their current intelligence and happiness, including comments such as "I'm quite content with my current self" and "I am the way I am for a reason." Those who would decrease their intelligence focused on the value of happiness for them, with comments such as "I value happiness far more," "Happy is important," and "Happiness is important to me." This suggests people may be treating happiness as a value in its own right, to which a utility can be assigned, rather than being a direct measure of utility. Those who preferred to increase their intelligence focused in part on the increase in opportunities and affordances, with comments such as "Being more intelligent affords me more opportunities," "I'll have more opportunities," "I could achieve more goals than I am now," "Intelligence can take me further than happiness," and "Intelligence has the capacity to help people." Thus, intelligence was seen mainly as an enabling quality rather than a value in and of itself, although one person did explain, "I care about knowledge."

[12] A possible objection here is that people are distinguishing "true" happiness from "fake" happiness. However, to make sense of true and false happiness already requires going beyond a naïve model, and we would suggest such a distinction is exactly captured by the machinery discussed in Equation (BiasedMetaAgent). That is, the current self imagines itself as a new, less intelligent self and finds such a situation abhorrent, even though the self that find this abhorrent would no longer exist if it were actually to undergo the transformation. "True" happiness in this sense is evaluated relative to the current deciding self.

4. General Discussion

L. A. Paul's (2014) recent philosophical inquiry presents two challenges to any formal account of how people can choose to undergo transformative experiences. The first challenge is that of novelty: A rational approach to decision-making requires us to imagine what it would be like to undergo an particular experience, but some experiences (both big and small) are outside our capability to do so. The second challenge is that of change: By transforming we become someone else, with potentially different views on whether we should have undergone that change. The two challenges are related but independent. Even if one knows completely what a new self will be like, there is still a rational difficulty in choosing to become (or not become) that self. In this chapter, we tried to meet both challenges using computational frameworks, which try to ground the everyday intuition that people have when thinking about change and new experience. Our models do not directly address how people ultimately make their decision, but rather how they conceptualize transformative decisions. This division is similar to the one that exists between any intuitive theory and the real world. While we use a theory-of-mind to predict and explain other people and their actions, their actual action-execution and decision-making process may work in a completely different way. Similarly, the way that we explain and predict our own actions may be unrelated to the way we actually make decisions (Saxe 2009).

Our answer to the challenge of novelty was a formal account demonstrating how people could reasonably choose to experience new things, by leveraging their own pre-existing structured view of the world. The model does not, and is not meant to, solve the problem of decisions about novel experiences. Rather, the model shows why some decisions about novel experience might be harder than others. When considering an experience that is a sub-category of a more general and well-understood structure (a yellow grape is a grape, which is a food, and so on), people can rely on their past experience, preference, and understanding. For new categories that are higher up in the hierarchy (such as trying a new type of food altogether), people may rely on their hyper-preferences over categories or novel experiences more generally. For decisions that are far outside the realm of understood experience, such as trading a current sense for a new one, our model would be hard pressed to find relevant preferences and experiences to draw on—and it is telling that these decisions also feel intuitively more challenging. Even in such cases, it is also potentially possible for people to leverage their past experiences with preferences and utility change ("I have experienced preference change in the past and things turned out well, so I should try it again").[13] Such leveraging is not captured by our model, though as with our model it would rely on identifying higher-order features of an experience, which our model can target for having a utility function or preference over.

[13] We thank John Schwenkler for this point.

On top of this, our work highlights the non-trivial nature of even a simple novel decisions, requiring tools from current non-parametric statistics. In this sense, philosophy and computational modelling share the purpose of showing the obvious to be non-obvious. People may think it trivial to reasonably choose to eat a new type of grape, but both a philosophical inquiry and a modelling attempt show it is not so. Such an obliviousness to the complexity of the unconscious computations supporting thought is not unique to this particular domain (consider how obvious it seems to hear and understand a friend's words, and then consider the computation that must go into that).

In response to the challenge of change, we built a framework of arbitration between possible selves. While much work needs to be done to further validate these models, empirical evidence does suggest that people do not adopt a strict "view-from-nowhen" when arbitrating between possible selves, and may treat future beliefs and utilities differently. This contrasts with any formalism of transformative decisions that rely upon "global" or perspective-independent decision functions (Pettigrew 2015) precisely insofar as agents are unable to treat those non-present beliefs and utilities as their own (Moran 2001).

The aim of our framework in the second half was to capture some of the flavor of thinking about transformative decision-making. In particular, our models address the tyranny of the current self over the potential future self. The tyranny of the present may extend into the past as well as the future: once a decision has been made to transform, the experiences of the past self are evaluated through the lens of the utilities and beliefs of the current, transformed self. Future work could explore models of agent's post hoc ratiocination about transformative experiences, and the ways in which it is comparable to ratiocination planning about transformative experiences in light of this tyranny of the present; our present selves did not have time for it.

While we hope our work is useful formally and empirically, much of our work is both suggestive and tentative. In particular, further work needs to be done to disentangle our claims about the reasoning of transformations from the means of those transformations, as well as how an agent can verify that a transformation did indeed take place. Our models also leave out several key features unique to big decisions, and it is worth considering the outlines of a formal framework that could in principle address those.

One such key feature is the long shadow cast by the choice not taken (Ullmann- Margali 2006). Unlike a transformative *experience,* which may or may not evoke feelings of regret, transformative *decisions* are riddled with counterfactual worries. Our modeling does not specifically address such potential regrets, in that they are no different from standard utility theories of rational choice under uncertainty, or from the more psychology-based explanations of prospect theory (Kahneman 2011). There are several proposals for formal models of decision-making that take regret into account (e.g. Bell 1982; Loomes and Sugden 1982); but while the causal role of regret is recognized as powerful and important, no current account is as accepted as utility theory or prospect

theory. Once such accounts are considered, we may ultimately find that *transformative* decision-making does not need any additional conceptual parts to address regret.

Another key feature of transformative decision-making is this: big decisions are hard to make. They are agonizing. People will go out of their way to avoid thinking about them at all, putting off the decision or whittling it down into manageable pieces (Ullmann-Margalit 2006). When people do bring themselves to think about such decisions, they may cycle through the possible futures, end in indecision, and then repeat the process later. The meta-agents in our framework encounter none of these difficulties or cycling behavior, and it is worth considering in general terms what formal model *could* account for this feature.

A computation may be difficult due to basic limitations of memory and time, such as when a search algorithm must traverse a rapidly fanning tree of options. Thus it may be that, due to the many options involved in imagining two different lives, the subjective analog of this computational difficulty is agony and frustration. But many decisions can be resource-draining in the sense that they involve many options, without being agonizing in the way transformative decisions are. Furthermore, a computationally difficult problem may mean we do want to spend more time and resources getting the answer right, while people actively *avoid* spending time thinking about big decisions. Finally, spending additional time revisiting a big decision does not seem to produce new imaginings of the future; we learn nothing more about our future selves, but rather go through a cycle of considering the same pros and cons over and over (although there is room for empirically testing this statement).

It may also be that there is no one unique reason accounting for the difficulty of big decisions. People in general dislike making decisions, experience a cost in cost– benefit analysis, shirk responsibility, and deploy second-order strategies to avoid deploying their decision-making apparatus (Bobadilla-Suarez etal. 2017; Sunstein and Ullmann-Margalit 1999). If this is the case, models of transformative experience would only have to adopt the amalgam of formal psychological models for the different aspects that make any decision difficult, and inflate them as necessary in terms of cost, responsibility, and the like.

Still, it seems to us that there is one important and distinct feature of the difficulty of transformative decisions, which may require a separate computational analog: Choosing to transform or not to transform forcefully severs the other entertained self. It may be that the state of having future options is in and of itself pleasurable, and cutting off a major branch of a potential time-line is thus painful to contemplate and carry out. Our imagined transformed self is not a completely alien being, but rather it is a part of how we see ourselves right now. Transformative decisions may be difficult insofar as they force us to let go of that part of ourselves. Torschlusspanik is painful in and of itself, but it may particularly painful when we are the ones shutting the gate.

For the technical appendix, please see: www.samuelzimmerman.com/ transformativeAppendix>.

References

Baker, C. L., J. Jara-Ettinger, R. Saxe, and J. B. Tenenbaum. 2017. "Rational Quantitative Attribution of Beliefs, Desires and Percepts in Human Mentalizing." Nature Human *Behavior* 4: 1-10.

Baker, C. L., R. Saxe, and J. B. Tenenbaum. 2009. "Action Understanding as Inverse Planning." *Cognition* 113(3): 329-49.

Baker, C. L., R. Saxe, and J. B. Tenenbaum. 2011. "Bayesian Theory of Mind: Modeling Joint Belief–Desire Attribution." *Proceedings of the Annual Meeting of the Cognitive Science Society 33.*

Baker, C. L., and J. B. Tenenbaum. 2014. "Modeling Human Plan Recognition Using Bayesian Theory of Mind." In G. Sukthankar, R. P. Goldman, C. Geib, D. Pynadath, and H. Bui (eds), *Plan, Activity, and Intent Recognition: Theory and Practice,* 177-98. Burlington, MA: Morgan Kaufmann.

Bell, D. E. 1982. "Regret in Decision-making under Uncertainty." *Operaions Research* 30(5): 961-81.

Bentham, J. 1996. *The Collected Works of Jeremy Bentham: An Introduction to the Principles of Morals and Legislation.* Oxford: Clarendon Press.

Bobadilla-Suarez, S., C. R. Sunstein, and T. Sharot. 2017. "The Intrinsic Value of Choice: The Propensity to Under-Delegate in the Face of Potential Gains and Losses." *Journal of Risk and Uncertainty* 54(3): 187-202.

Dennett, D. C. 1989. *The Intentional Stance.* Cambridge, MA: MIT Press.

Elster, J., and G. Loewenstein. 1992. "Utility from Memory and Anticipation." In G. Loewenstein and J. Elster (eds), *Choice Over Time,* 213-34. New York: Russell Sage Foundation.

Gershman, S. J., J. Malamud, and J. B. Tenenbaum. 2017. "Structured Representations of Utility in Combinatorial Domains." *Decision* 4(2): 67-86.

Goodman, N. D., and J. B. Tenenbaum. n.d. *Probabilistic Models of Cognition,* 2nd edn: <`https://www.probmods.org/`>

Gopnik, A. 1993. "Theories and Illusions." *Behavioral and Brain Sciences* 16(1): 90-100.

Griffiths, T. L., C. Kemp, and J. B. Tenenbaum. 2008. "Bayesian Models of Cognition." In

R. Sun (ed.), *The Cambridge Handbook of Computational Psychology,* 59-100. Cambridge: Cambridge University Press.

Hamlin, K. J. 2013. "Moral Judgment and Action in Preverbial Infnts and Toddlers: Evidence for an Innate Moral Core." *Current Directions in Psychological Science* 22: 186-93.

Happe, F. 2003. "Theory of Mind and Self." *Annals of the New York Academy of Sciences* 1001(1): 134-44.

Jackson, F. 1986. "What Mary Didn't Know." *Journal of Philosophy* 83(5): 291-5.

Jara-Ettinger, J., H. Gweon, L. E. Schulz, and J. B. Tenenbaum. 2015. "Children's Understanding of the Costs and Rewards Underlying Rational Action." *Cognition* 140: 14-23.

Jara-Ettinger, J., H. Gweon, L. E. Schulz, and J. B. Tenenbaum. 2016. "The Naive Utility Calculus: Computational Principles Underlying Commonsense Psychology." *Trends in Cognitive Sciences* 20: 1-16.

Jern, A., and C. Kemp. 2015. "A Decision Network Account of Reasoning About Other People's Choices." *Cognition* 142: 12-38.

Kaelbling, L. P., M. L. Littman, and A. R. Cassandra. 1998. "Planning and Acting in Partially Observable Stochastic Domains." *Artificial Intelligence* 101(1): 99-134.

Kahneman, D. 2011. *Thinking, Fast and Slow.* London: Macmillan.

Kahneman, D., etal. 2003. "Experienced Utility and Objective Happiness: A Moment-Based Approach." *Psychology of Economic Decisions* 1: 187-208.

Kahneman, D., P. P. Wakker, and R. Sarin. 1997. "Back to Bentham? Explorations of Experienced Utility." *Quarterly Journal of Economics* 112(2): 375-406.

Kemp, C., A. Perfors, and J. B. Tenenbaum. 2007. "Learning Overhypotheses with Hierarchical Bayesian Models." *Developmental Science* 10(3): 307-21.

Lewis, D. 1990. "What Experience Teaches." In W. Lycan (ed.), *Mind and Cognition: A Reader*, 499-519. Oxford: Blackwell.

Locke, J. 1700. *An Essay Concerning Human Understanding.* Awnsham and John Churchil, at the Black-Swan in Pater-Noster-Row, and Samuel Manship, at the Ship in Cornhill, near the Royal-Exchange.

Loomes, G., and R. Sugden. 1982. "Regret Theory: An Alternative Theory of Rational Choice under Uncertainty." *Economic Journal* 92(368): 805-24.

McCrae, R. R., and O. P. John. 1992. "An Introduction to the Five-Factor Model and Its Applications." *Jounal of Personality* 60(2): 175-215.

Moran, R. 2001. *Authority and Estrangement: An Essay on Self-Knowledge.* Princeton, NJ: Princeton University Press.

Nagel, T. 1989. *The View From Nowhere.* Oxford: Oxford University Press.

Newman, G. E., P. Bloom, and J. Knobe. 2014. "Value Judgments and the True Self." *Personality and Social Psychology Bulletin* 40(2): 203-16.

Parfit, D. 1984. *Reasons and Persons.* Oxford: Oxford University Press.

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

Paul, L. A. 2015a. "What You Can't Expect When You're Expecting." *Res Philosophica* 92(2): 149-70.

Paul, L. A. 2015b. "Transformative Experience: Replies to Pettigrew, Barnes and Campbell." *Philosophy and Phenomenological Research* 91(3): 794-813.

Perner, J., S. R. Leekum, and H. Wimmer. 1987. "Three-Year-Olds' Difficulty With False Belief: The Case for a Conceptual Deficit." *British Journal of Developmental Psychology* 5(2): 125-37.

Pettigrew, R. 2015. "Transformative Experience and Decision Theory." *Philosophy and Phenomenological Research* 91(3): 766-74.

Robinson, T. 1987. *Heraclitus: Fragments.* Toronto: University of Toronto Press.

Russell, S., and P. Norvig. 1995. *Artificial Intelligence: A Modern Approach.* Upper Saddle River, NJ: Prentice Hall.

Saxe, R. 2009. "The Happiness of the Fish: Evidence for a Common Theory of One's Own and Others' Actions." In K. D. Markman, W. M. P. Klein, and J. A. Suhr (eds), *Handbook of Imagination and Mental Simulation,* 257-65. Brighton: Psychology Press.

Shoemaker, S. 1995. "Moore's Paradox and Self-Knowledge." *Philosophical Studies* 77(2): 211-28.

Strohminger, N., J. Knobe, and G. Newman. 2017. "The True Self: A Psychological Concept Distinct From the Self." *Perspectives on Psychological Science* 12(4): 551-60.

Strohminger, N., and S. Nichols. 2014. "The Essential Moral Self." *Cognition* 131(1): 159-71.

Sunstein, C. R., and E. Ullmann-Margarit. 1999. Second-Order Decisions. *Ethics* 110(1): 5-31.

Tenenbaum, J. B., C. Kemp, T. L. Griffiths, and N. D. Goodman. 2011. "How to Grow a Mind: Statistics, Structure, and Abstraction." *Science* 331: 1279-85.

Thaler, R. H., and C. R. Sunstein. 2008. *Nudge: Improving Decisions About Health, Wealth, and Happiness.* New Haven, CT: Yale University Press.

Todorov, E. 2004. "Optimality Principles in Sensorimotor Control." *Nature Neuroscience* 7(9): 907-15.

Ullmann-Margalit, E. 2006. "Big Decisions: Opting, Converting, Drifting." In A. O'Hear (ed.), *Political Philosophy,* 157-72. Cambridge: Cambridge University Press.

Weirich, P. (2004). *Realistic Decision Theory: Rules for Nonideal Agents in Nonideal Circumstances.* Oxford: Oxford University Press.

5. Transformative Experience and the Knowledge Norms for Action: Moss on Paul's Challenge to Decision Theory[7]

Richard Pettigrew[8]

1. Introduction

Experience can change you in many different ways. It can teach you what it's like to have a particular new experience—for instance, the experience of eating oysters for the first time, the experience of understanding a complex mathematical proof for the first time, or the experience of being a parent for the first time. If an experience changes you in this way, L. A. Paul dubs it an epistemically transformative experience (henceforth, ETE). But experience can also change "who you are" by changing some of your deepest values or the convictions that form the foundation of your system of beliefs—for instance, the experience of holding your newborn baby in your arms for the first time might change the intensity with which you value family; the experience of awe-inspiring beauty in the natural world might change your religious convictions, and with them many of your values; and the experience of spending time with people with

Down's syndrome might change the value you attach to their lives. If it changes you in this second way, it is a *personally transformative experience* (Paul 2014; 2015b).

Paul argues that the possibility of both sorts of experience poses serious and novel problems for the orthodox theory of rational choice, namely, expected utility theory.[1] In this chapter, I will focus only on Paul's argument that the possibility of ETEs raises

[1] While Paul targets expected utility theory with her objections, if they work, they affect the whole gamut of non-expected utility theories as well (Kahneman and Tversky, 1979; Quiggin, 1993; Buchak, 2013; Wakker, 2010).

[7] Thanks to Enoch Lambert, Sarah Moss, Laurie Paul, and John Schwenkler for their insightful and generous comments on earlier drafts of this chapter.

[8] Richard Pettigrew, Transformative Experience and the Knowledge Norms for Action: Moss on

a challenge for expected utility theory—I will call her objection *the Utility Ignorance Objection.* In a pair of earlier papers, I responded to Paul's challenge (Pettigrew 2015; 2016), and a number of other philosophers have responded in similar ways (Dougherty etal. 2015; Harman 2015)—I will call the argument that we have each put forward *the Fine-Graining Response.* Paul (2014) has her own reply to this response, which we might call *the Authenticity Reply.* But Sarah Moss has offered an alternative reply to the Fine-Graining Response on Paul's behalf (Moss 2018)— we'll call it *the No Knowledge Reply.* This appeals to the knowledge norms for action, together with Moss's novel and intriguing account of probabilistic knowledge. In this chapter, I consider Moss's reply and argue that it fails. I argue first that it fails as a reply made on Paul's behalf, since it forces us to abandon many of the features of Paul's challenge that make it distinctive and with which Paul herself is particularly concerned. Then I argue that it fails as a reply independent of its fidelity to Paul's intentions.

Before we can state Paul's challenge to decision theory, we have to make clear exactly which version of that theory she wishes to challenge (Section 2). Having done that, we can state Paul's Utility Ignorance Objection to that theory (Section 3), the Fine-Graining Response (Section 4), and Paul's Authenticity Reply (Section 5). Then, we introduce Moss's No Knowledge Reply (Section 6). We argue that it fails both as a reply on behalf of Paul (Section 7) and as a reply on its own (Sec8).

2. What Is Decision Theory?

Paul addresses her Utility Ignorance Objection to a particular interpretation of decision theory, namely, the realist-deliberative interpretation (though she sometimes hints that the objection is intended to apply to the realist-evaluative interpretation as well). Let me introduce these now by drawing two distinctions: the realist/ constructivist distinction and the deliberative/evaluative distinction.

First, the realist/constructivist distinction. Realist and constructivist understandings of decision theory agree on the building blocks of decision theory; they differ on which of these building blocks are more fundamental. On both, we have the following:

- a set of possible actions the agent might perform—call that set A;
- a set of possible states of the world—call it S;
- a preference ordering $\boxtimes$ over the possible actions in A;

Paul's Challenge to Decision Theory In: Becoming Someone New: Essays on Transformative Experience, Choice, and Change. Edited by: Enoch Lambert and John Schwenkler, Oxford University Press (2020).
© Richard Pettigrew.
DOI: 10.1093/oso/9780198823735.003.0006

- a credence function P over combinations of actions from A and states from A— given a in A and s in S, <verbatim>P(s	a)</verbatim> records how strongly the agent believes that the world is in state s under the supposition that they perform a;[a] and

[a] If the supposition in question is indicative supposition, the resulting theory is *evidential decision theory* (Jeffrey 1983); if it is subjunctive supposition, the result is *causal decision theory* (Joyce 1999). The difference between them will not matter here.

- a utility function U over combinations of actions from A and states from S— given a in A and s in S, *U(a,s)* records how strongly the agent values or desires or wants or endorses the outcome of performing action a when the world is in state s.

According to the constructivist understanding of decision theory, the preference ordering is primary—for some constructivists of a more behaviorist persuasion, only the preference ordering is psychologically real; for others, credence and utility functions are real too, but the preference ordering is more fundamental than both. Typically, constructivists then show that, providing that the preference ordering $\boxtimes$, set of states S, and the set of possible actions S satisfy particular structural constraints—the Savage (1954), Jeffrey (1983), or Joyce (1999) axioms, for instance—there is a credence function P and a utility function U such that the preference ordering places one action at least as high as another iff the expected utility of the first is at least as great as the expected utility of the second; that is, $a1 < a_2$ iff $ExpP(U(\text{a1})) < ExpP(U(\text{a}_2))$, where $ExpP(U(\text{a})) = \mathbf{X}\ \text{P(s||a)U(a,s)}$.[s2S]

That is, the preference ordering, which is primary, can be represented by the credence function and the utility function, which are thereby secondary. Such a result is known as a representation theorem, and they are central to the constructivist understanding of decision theory. On the other hand, the realist understanding of decision theory says that the credence function and utility function are primary, while the preference ordering on actions given by their expected utility relative to that credence function and utility function is secondary. Thus, for the constructivist, we begin with the preference ordering and derive the credence function and utility function; for the realist, we begin with the credence function and utility function and derive the preference ordering. For both, faced with a decision between a range of possible actions, an agent is rationally required to choose an action that sits at the top of the preference ordering—that is, it is irrational to choose an action a in A when there is an alternative action a* in A such that a ^ a*. Paul's challenge applies primarily to the realist understanding of decision theory.

Next, let's turn to the deliberative/evaluative distinction. Deliberative and evaluative understandings of decision theory differ on which elements of a decision are relevant to its rationality. For those who favor a deliberative understanding, decision

theory governs not only the choice that an agent makes in a given situation, but also the deliberation by which she comes to make that choice. In contrast, those who favor an evaluative understanding say that decision theory evaluates the choice only. Thus, for instance, suppose I must decide whether or not to take an umbrella when I leave my house. As it happens, I would maximize my expected utility by taking the umbrella—I think it's pretty likely to rain, I hate getting wet, and it doesn't much bother me to carry the umbrella. Now suppose that I do indeed end up taking the umbrella. But suppose that my reason for doing so was not that it would maximize my expected utility—it was not by calculating which action would maximize expected utility and then picking it that I reasoned to my conclusion. Rather, I chose the action I did simply using the rule: Always pick the action that involves approximating most closely the sartorial choices of Mary Poppins. Then, according to the evaluative understanding of decision theory, I am fully rational, because I chose the option that maximizes expected utility, while according to the deliberative understanding, I am not, because I did not deliberate correctly concerning my choice—my decision was not sensitive to the expected utility of the actions between which I had to choose. As we will see below, Paul's challenge applies primarily to the deliberative understanding of decision theory, though I will also ask whether Paul's insight supports a stronger argument, which tells against the evaluative understanding as well.

3. Paul's Utility Ignorance Objection

What is an ETE? It is an experience that teaches you something that you couldn't come to know without having that experience. Thus, for Frank Jackson's scientist, Mary, who has lived her whole life in a monochrome black-and-white room, the experience of stepping outside and seeing the color red for the first time is an ETE (Jackson 1986). However much Mary learned about the physical properties of red objects, she could not know what it is like to see red. Similarly, for some people, becoming a parent for the first time is an ETE. However much they attended to the testimony of people who already have children, however much they read novels about parenting, they couldn't know what it was going to be like to be a parent until they became one themselves (Paul 2014).

In Mary's case, what she learns from her ETE is a phenomenological fact—she learns what it is like to see red. In the case of the new parent, there is likely a phenomenological component to what they learn from the experience as well—they learn what it is like to feel a particular sort of bond with another person; and they learn what it is like to have responsibility for another life. But there may well be other components—the experience might teach you some moral facts, for instance. For Paul's objection, she needs only this: ETEs teach you something that you cannot learn any other way and that you need to know in order to know the utility that you assign to the outcomes of certain actions that are available to you.

For instance, suppose I must decide whether or not to apply to adopt a child and become a parent—this will serve as our illustrative example throughout the chapter. If I choose to apply and my application is successful, I will become a parent. In order to calculate the expected value of choosing to apply, I must therefore know the utility I assign to the outcome on which I apply and my application is successful. But in order to know that, I need to know what it will be like to be a parent. And that, for some people, is something that they can know only once they become a parent. For such people, then, it seems that the ingredients that they require in order to calculate their expected utility for applying to adopt a child are not epistemically available to them. And thus they are barred from deliberating in the way that the realist-deliberative understanding of decision theory requires of them. They are unable to make the decision rationally.

Using the ingredients of decision theory introduced above, we can state the problem as follows: there are two actions between which I must choose—apply to adopt a child (Apply); don't apply (Don't Apply). And let's say that there are two states of the world— one in which I would be successful if I were to apply (Succeed), and one in which I would be unsuccessful (Fail). So the decision table is as shown in Table 5.I.[2]

According to the realist, to choose whether or not to apply, I must determine whether I prefer applying to not applying—that is, whether *Apply ^ Dont Apply* or *Apply ^ Dont Apply.* To determine that, I must calculate the expected utility of those two actions relative to my credence function and my utility function. And to calculate that, I must know what my credence is in each of the two possible states of the world under the supposition of each of the possible actions—that is, I must know *P(Succeed* jj *Apply)* and *P(Fail* jj *Apply),* as well as *P(Succeed* jj *Don't Apply)* and *P(Fail* jj *Don t Apply).* And I must know my utilities for the different possible outcomes—that is, I must know *U (Apply, Succeed)* and *U (Apply, Fail)* and *U (Don't Apply, Succeed)* and *U (Don't Apply, Fail).* The problem that Paul identifies is that it is impossible to know *U(Apply, Succeed)* prior to making the decision and becoming a parent; and thus it is impossible to deliberate about the decision in the way that the realist-deliberative understanding of decision theory requires. Paul concludes that there is no rational way to make such decisions. This is the Utility Ignorance Objection to the realist-deliberative understanding of decision theory.

Table 5.1 Our decision table

	Succeed	***Fail***
Apply	???U(Apply, Succeed)???	U(Apply, Fail)
Don't Apply	U(Don't Apply, Succeed)	U(Don't Apply, Fail)

[2] In fact, you might go further and say that applying and failing in your application is also an epistemically transformative experience, given your emotional investment in the application. If that's the case, there should be question marks around the top-right entry in our table, and Paul's problem is even more pronounced.

Before we move on to consider how we might respond to this objection, let us pause a moment to consider its scope. First, note that the challenge targets only the realist understanding of decision theory, not the constructivist. For the constructivist, my credence and utility functions are determined by my preference ordering. Thus, to know them I need only know my preference ordering. And for many constructivists I can know that simply by observing how I choose between given sets of actions—I prefer a to b just in case I would choose a over b in a binary choice between them. Paul's challenge applies only when we take the utility function to determine, at least in part, our preference ordering, as the realist does. Second, note that, at least primarily, the challenge targets only the deliberative understanding of decision theory, not the evaluative. On the realist-evaluative understanding, I do not need to know my credences or my utility function in order to be rational. On this understanding, in order to be rational, I need only choose the action that *in fact* maximizes expected utility; I need not choose it *because* it maximizes expected utility. Thus, Paul's argument has no bite for the evaluative understanding.

Now, we might try to extend Paul's argument so that it does apply to the realist-evaluative understanding. To do that, we need to argue not only that I do not *know* the utility *U(Apply, Succeed)* prior to my choice between *Apply* and *Don't Apply*, but indeed that *U(Apply, Succeed)* is not even *determined* prior to that choice—and indeed Paul herself hints at that interpretation in some places (Paul 2015a: 494). If that were the case, then there would be no way to make the choice rationally, even according to the realist-evaluative understanding. I am skeptical that the argument as it stands can support this conclusion. The examples that Paul gives motivate the claim that we cannot know our utilities for certain outcomes. To move from that to the claim that those utilities are not determined requires a particular sort of account of what they are. For instance, if we are hedonists, and the utility we assign to an outcome measures the intensity of the pleasure that we would experience in that outcome, then it is quite possible that those utilities might be determined but unknowable. In order to extend Paul's argument, then, we must at least rule out such a conception of utilities, and that requires further argument.

4. The Fine-Graining Response

There is a natural response to Paul's challenge. Expected utility theory is designed to deal with decisions made in the face of uncertainty. Usually that uncertainty concerns the way the world is beyond or outside of the agent. For instance, suppose I'm uncertain whether my adoption application would be successful if I were to apply. Then, when I'm making my decision, I ensure that the set of possible states of the world includes one in which my application succeeds and one in which it fails. I then quantify my uncertainty concerning these two possibilities in my credence function, and I use that to calculate my expected utility—perhaps I know that only 12% of adoption applications succeed,

and I set my credence that mine will succeed to 0.12 in line with that. However, there is no reason why the uncertainty quantified by my credence function should concern only the way the world is beyond me. What Paul's argument shows is that I am uncertain not only about the world, but also about the utility that I assign to becoming a parent; I am uncertain not only about whether Succeed or Fail is true, but also about the value U(Apply, Succeed). Thus, just as I ensured that my decision problem includes possible states of the world at which I succeed in my application and possible states where I fail, similarly I should respond to Paul's challenge by ensuring that my decision problem includes possible states of the world at which I become a parent and value it greatly, possible states at which I become a parent and value it a moderate amount, states at which I become a parent and value it very little, and so on. Having done this, I should quantify my uncertainty concerning the utility I assign to being a parent in my credence function, and use that to calculate my expected utility as before.

More precisely, and simplifying greatly, suppose the possible utility values that I might assign to being a parent are —12, 3, and 10. Then, while my original set of possible states of the world is S = *{Succeed, Fail},* my new expanded set of possible states of the world is

S* = {Succeed & utility of being a parent is —12*,* Succeed & utility of being a parent is 3*,* Succeed & utility of being a parent is 10*, Fail}.*

Now, recall the problem that Paul identified. Given the original way of setting up the decision problem, in order to deliberate rationally between *Apply* and *Don't Apply*, I need to know the utilities I assign to each possible outcome of each of the possible actions. In particular, I need to know *U(Apply, Succeed).* But I can't know that until I make the decision and become a parent. However, on the new formulation of the decision problem, with the expanded set of states S^l, I do know the utilities I assign to each possible outcome of each of the possible actions. For I know that:

- U (Apply, Succeed & utility of being a parent is -12)

= 12,

- U (Apply, Succeed & utility of being a parent is 3)

= 3,

- U (Apply, Succeed & utility of being a parent is 10)

= 10.

Next, I quantify my uncertainty in these new possible states to give:

P(Succeed & utility of being a parent is -12 || *a),*

P(Succeed & utility of being a parent is 3 || *a),*

P(Succeed & utility of being a parent is 10 || *a),*

P(Fail || *a*).

where a is either *Apply* or *Don't Apply*. And, given this, I can calculate my expected utility and discharge the obligations of rationality imposed by the realist-deliberative understanding of decision theory. Paul's Utility Ignorance Objection, it seems, is answered. Call this the Fine-Graining Response, since it involves expanding, or fine-graining, the set of possible states of the world.

5. Paul's Authenticity Reply

Paul herself is not satisfied with this response. She allows that I can expand the set of possible states of the world in the way described. And she allows that I can form credences in those different states of the world. But she worries about the sort of evidence on which I might base those credences.

Let's start with an ordinary decision that does not involve an ETE. Suppose I am deciding whether to have chocolate ice cream or strawberry ice cream. I have tasted both in the past, so I know what both experiences will be like—neither experience would be transformative. As a result, when I come to make my decision, I know the utility I assign to the outcome in which I eat chocolate ice cream. I know it by imaginatively projecting myself forward into the situation in which I am eating chocolate ice cream. And I can do this because I have tasted chocolate ice cream in the past. And similarly for the utility I assign to the outcome in which I eat strawberry ice cream. I know what it is, and I know it because I've tasted strawberry ice cream in the past and so I can imaginatively project myself forward into the situation in which I'm eating it.

When I consider the utility I assign to becoming a parent, I can't imaginatively project in this way, since I'm not a parent and becoming a parent is an ETE. As described above, I respond to this epistemic barrier by expanding the set of possible states of the world I consider in my decision problem. I expand them so that they are fine-grained enough that each specifies my utility for becoming a parent at that world; and my credences in these different possible states quantify my uncertainty over them. But how do I set those credences? I cannot do anything akin to imaginatively projecting myself into the situation of being a parent, as I did with the chocolate ice cream, because becoming a parent is an ETE. What can I do instead?

Well, the natural thing to do is to seek out the testimony of people who have already undergone that transformative experience.[3] Perhaps I cannot discover from them exactly what it is like to be a parent—since it's an ETE, the only way to learn what it's like is to undergo the experience. But perhaps I can learn from them how much they value the experience. And after all, that's all that I need to know in order to make my decision rationally, according to the realist-deliberative understanding of decision theory—expected utility theory doesn't require that you know what an outcome will be like; it requires only that you know how much you value it and thus

[3] See Dougherty etal. (2015), for two further ways in which I might set these credences. I focus on testimonial evidence here since it is the sort of evidence that Moss considers.

how much it contributes to the expected utility calculation. However, as we all know, different people value being a parent differently. For some, it is an experience of greater value than all other experiences they have in their life. For others, it is a positive experience, but doesn't surpass the value of reciprocated romantic love, or extremely close friendships, or succeeding in a career, or helping others. And for yet others, it is a negative experience, one that they would rather not have had. Simplifying greatly once again, let's assume that all parents fall into these three groups: members of the first assign 10 utiles to the outcome in which they become a parent; members of the second assign 3; and members of the third assign -12. And let's assume that 10% fall into the first group; 60% into the second; and 30% into the third. Now, suppose that I learn this statistical fact by attending to the testimony of parents. Then I might set my credences as follows (where we assume for convenience that I am certain that my adoption application will be successful, so *P(Succeed || Apply)* = 1):

- P(Succeed & utility of being a parent is -12 <verbatim>	</verbatim>

a) = 0.3,

- P(Succeed & utility of being a parent is 3 <verbatim>	</verbatim>

a) = 0.6,

- P(Succeed & utility of being a parent is 10 <verbatim>	</verbatim>

a) = 0.1.)

With these in hand, I can then calculate the expected utility of *Apply* and *Don't Apply,* I can compare them, and I can make the choice between them in the way that the realist-deliberative decision theorist requires.

However, Paul claims that if I choose in this way then my decision is badly flawed. She holds that an agent who made the decision to become a parent in this way would be "alienated" from that decision; the choice thus made would be "inauthentic":

> A [...] problem with leaving your subjective perspective out of your decisions connects to the Sartrean point that making choices authentically and responsibly requires you to make them from your first personal perspective. A way to put this is that if we eliminate the first personal perspective from our choice, we give up on authentically owning the decision, because we give up on making the decisions for ourselves. We give up our authenticity

> if we don't take our own reasons, values, and motives into account when we choose. To be forced to give up the first person perspective in order to be rational would mean that we were forced to engage in a form of self-denial in order to be rational agents. We would face a future determined by Big Data or Big Morality rather than by personal deliberation and authentic choice.
>
> (Paul 2014: 130)

For Paul, then, the problem lies in the way that I set my credences in the fine-grained states of the world. I set my credences concerning my own utilities by deferring to statistical facts about how others assign different utilities. My evidence does not sufficiently concern *my* utilities; and thus I am alienated from any decision based on the credences that I form in response to that evidence. I am like the agent who makes a moral decision by deferring to societal norms or the value judgements of the majority group, rather than making those decisions herself. Paul contrasts this statistical method of forming opinions about my own utilities with the method described above in the case of the chocolate and strawberry ice cream, where I imaginatively project myself into the situation in which I have the experience based on my own memory of previous similar experiences. In those cases, the opinions formed do not give rise to the same sort of alienation and inauthenticity, since they are connected in the right way to my own utilities. They are more akin to the agent who makes the moral decision for themselves.

I have responded to Paul's concern elsewhere, where I argue that there is a crucial difference between these cases (Pettigrew 2015: 770). When I set my credences concerning my own utilities by appealing to the statistical evidence concerning the utilities of others, I do so because I think that this statistical evidence tells me something about my own utility; it is good evidence concerning my own utilities. In contrast, when I defer to societal norms to make a moral decision, I do so not because I think that those norms tell me anything about my own values; I do not think they provide good evidence concerning what I think is the correct moral action. I do so because I can't decide what I think is the correct moral action, or I do not have the courage to follow my own moral compass. I mention Paul's Authenticity Reply here partly for the sake of completeness, but also because Moss's No Knowledge Reply to the Fine-Graining Response also argues that the problem with such decisions lies in the nature of the evidence on the basis of which I form my credences about my utilities. Let's turn to Moss's reply now.

6. Moss's No Knowledge Reply

Suppose I set my credences in *Succeed & utility of being a parent is* -12, etc. (as in Section 5). That is, I set them on the basis of statistical evidence concerning the

utilities that existing parents assign to being a parent. For Paul, the problem is that such evidence does not sufficiently concern my utilities in particular; it is too much concerned with the utilities of other people. For Moss, the problem with those credences is not that they are not sufficiently concerned with me, or at least that is not the primary problem. Rather, the problem is that those credences do not constitute knowledge, and rational decisions must be based on credences that constitute knowledge (Moss 2018: sec. 9.5).

To those unfamiliar with Moss's work, it might sound as if she is making a category mistake. Credences, you might think, are simply not the sort of thing that can constitute knowledge. Full beliefs can—if I believe that it's raining, then that belief might count as knowledge. But credences, or partial beliefs, cannot—if I have credence 0.6 that it's raining, then it makes no more sense to say that that credence counts as knowledge than it does to say that a colorless idea sleeps furiously. Or so you might think. But Moss denies this (Moss 2013; 2018). Let's see why.

First, it is worth saying what Moss takes credences to be. Suppose I say that I'm 50% confident that Kenny is in Hamburg. On the standard interpretation, this means that I have a precise graded attitude—a credence—towards the standard, non- probabilistic content *Kenny is in Hamburg*, where the latter might be represented by a set of possible worlds. In particular, I have a 0.5 credence in that non-probabilistic content. For Moss, in contrast, a credence is not a graded attitude towards a standard propositional content; rather, it is a categorical attitude towards what she calls a *probabilistic content.* For instance, to say that I'm 50% confident that Kenny is in Hamburg is to say that I have a categorical attitude—in fact, a belief—towards the probabilistic content *Kenny is 50% likely to be in Hamburg.*

What are these probabilistic contents? Well, just as a standard propositional content, such as *Kenny is in Hamburg*, can be represented by a set of possible worlds, so a Mossian probabilistic content, such as *Kenny is 50% likely to be in Hamburg,* is represented by a set of probability spaces, where a probability space is a set of possible worlds together with a probability distribution defined over those worlds. Thus, the probabilistic content *Kenny is 50% likely to be in Hamburg* is represented by the set of those probability spaces in which the probability distribution assigns 50% to the proposition *Kenny is in Hamburg*—that is, the set {P: *P(Kenny is in Hamburg)* = 0.5}.

Another example: Suppose I say that I'm more confident than not that Kenny is in Hamburg. On the standard interpretation, this means that I have an imprecise graded attitude towards the propositional content *Kenny is in Hamburg.* Imprecise graded attitudes are also represented by sets of probability spaces—these are usually called *representors.* In this case, my imprecise graded attitude is represented by the set of those probability spaces in which the probability distribution assigns more than 50% to the proposition *Kenny is in Hamburg*—that is, the set {P: *P(Kenny is in Hamburg)* > 0.5}. That set is my representor. For Moss, in contrast, I do not have a graded attitude towards the propositional content *Kenny is in Hamburg*, but rather a

categorical attitude towards the probabilistic content *Kenny is more likely than not to be in Hamburg*. The probabilistic content towards which I have that categorical attitude is in fact represented by the same set of probability spaces that is used to represent the imprecise graded attitude that is usually attributed to me—that is, my representor, {P: *P(Kenny is in Hamburg)* > 0.5}.

Now, citing a large body of examples, Moss argues that we often say that, just as beliefs in standard, non-probabilistic contents—viz. propositions—can count as knowledge, so can beliefs in probabilistic contents—viz. the contents represented by sets of propositions. For instance, I might say that Patricia knows that Kenny is 50% likely to be in Hamburg, or that Jason knows that Kenny is more likely than not to be in Hamburg.

As well as citing intuitive examples in which we ascribe probabilistic knowledge, Moss also gives examples that show that there are distinctions between categorical beliefs in probabilistic contents that are analogous to the distinctions that we mark between different categorical beliefs in propositions by categorizing one as merely justified and the other as knowledge. For instance, suppose that I know that the objective chance of this coin landing heads is 60%. And my credence that it will land heads is 0.6—that is, in Moss's framework, I believe that the coin is 60% likely to land heads. Next, suppose that you also set your credence in heads to 0.6—that is, you also believe the coin is 60% likely to land heads. But you set your credence in this way not because you know the objective chance, but because you know that Sarah's credence in heads is 0.6 and you have good reason to take Sarah to be an expert on the bias of coins. However, while you are right that Sarah is generally expert on such matters, in this case she hasn't actually inspected the coin and instead just plucked a number from thin air.

In such a case, it seems that, while both of us have justified credences that are correct in a certain sense, yours is merely justified, while mine counts as knowledge.

Moss furnishes us with a splendidly detailed account of probabilistic knowledge, which includes a Bayesian expressivist semantics for probabilistic knowledge ascriptions as well as an account of the factivity, safety, and sensitivity conditions on probabilistic knowledge. But her No Knowledge Reply to the Fine-Graining Response does not depend on the more sophisticated or radical elements of her account. Rather, it depends on just three claims about probabilistic knowledge.

The first, we have met already: it is the claim that credences—and, more generally, beliefs in probabilistic contents—can count as knowledge, just as beliefs in non-probabilistic contents can.

The second claim concerns a certain sort of case in which the credences you form don't count as knowledge. Suppose we meet. Noting that I am a living human being, and knowing that about 0.7% of living human beings will die in the next year, you form a credence of 0.007 that I will die in the next year. Then, for Moss, your credence does not count as knowledge. The problem is that you cannot rule out relevant alternative reference classes to which I belong and amongst which the frequency of death within

the next year is quite different. For instance, you know that I am 35 years old. And you can't rule out that the likelihood of death amongst living 35-year-olds is quite different from the likelihood amongst all human beings. You know that I am male. And you can't rule out that the likelihood of death amongst living males is different from the likelihood amongst all human beings. And so on. You believe that it's 0.7% likely that I will die in the coming year, but you can't rule out that my death is X% likely, for a range of alternative values of X. Moss likens the case to Goldman's fake barn scenario (Goldman 1976). I am travelling through Fake Barn County, and I stop in front of a wooden structure that looks like a barn. I form the belief that the structure in front of me is a barn because that's what it looks like. But my visual experience cannot distinguish a barn from a barn facade. So I cannot rule out the alternative possibility that the structure is a barn facade. And this alternative is relevant because Fake Barn County lives up to its name: it's full of fake barns. Therefore, my belief cannot count as knowledge. Similarly, since you cannot rule certain alternative reference classes amongst which my likelihood of death within the next year is quite different from 0.7%, your credence of 0.007 that I will die in the next year cannot count as knowledge. Or so Moss says.

Now, recall our response outlined above to Paul's Utility Ignorance Objection to decision theory. Since I cannot know the utility I assign to being a parent, I expanded the set of possible states of the world so that, in each, my utility is specified; and then I quantified my uncertainty concerning these different utilities in my credences. Since I could not set those credences by imaginatively projecting myself into the position of being a parent, I had to set them by appealing to the statistical evidence concerning the utilities that existing parents assigned to being parents. Since the evidence for my credences is statistical, if it is to count as knowledge, I must be able to rule out relevant alternative reference classes to which I belong on which the statistics are quite different. For instance, suppose I set my credences in the different possible utilities by appealing to the statistics amongst all existing parents. Then there are certainly relevant alternative references classes that I should consider: the class of all male parents; the class of all gay male parents; the class of adoptive parents; the class of all parents with family and social support network similar to mine; and so on. Given the evidence on which I based my credences, I cannot rule out the possibility that the distribution of the three candidate utilities for being a parent is different in these reference classes from the distribution in the reference class on which I based my credences. Thus, according to Moss, my credences cannot count as knowledge.

Finally, the third claim upon which Moss bases her No Knowledge Reply to the Fine-Graining Response is a conjunction of a probabilistic knowledge norm for reasons and a probabilistic knowledge norm for decision—together, we refer to these as the Probabilistic Knowledge Norms for Action, following Moss.

Probabilistic Knowledge Norm for Reasons Your credal state can only provide a reason for a particular choice if it counts as knowledge.

Probabilistic Knowledge Norm for Decisions Suppose the strongest probabilistic content you know is represented by a set **P** of probability functions; and suppose you are faced with a choice between a range of options. It is permissible for you to choose a particular option iff that option is permissible, according to the correct decision theory for imprecise credences, for an agent whose imprecise credal state is represented by **P**.

For instance, suppose you must choose whether to take an umbrella with you when you leave the house. The strongest proposition you know is represented by the set of probability spaces, $P = \{c\text{: } 0.4 < P(Rain) < 0.9\text{g}$. If rain is 90% likely, then taking the umbrella maximizes expected utility; if it is only 40% likely, then leaving the umbrella maximizes expected utility. Now imagine an agent whose credal state is represented by $\boldsymbol{P}$—in the language introduced above, $\boldsymbol{P}$ is her representor.[4] Which actions are permissible for this agent? According to some decision theories for imprecise credences, an action is permissible iff it maximizes expected utility relative to *at least one member of the representor*. We might call these *liberal* decision theories, since they make many actions permissible. On this decision theory, it is permissible to take the umbrella and permissible to leave it. Thus, according to the Probabilistic Knowledge Norm for Decisions, both actions are also permissible. According to other decision theories, an action is permissible iff it maximizes expected utility relative to *all members of the representor*. We might call these *conservative* decision theories, since they make few actions permissible. On this decision theory, neither taking nor leaving the umbrella is permissible for the agent with representor $\boldsymbol{P}$, and thus, according to the Probabilistic Knowledge Norm for Decisions, neither is permissible for me.

Thus, putting together the various components of Moss's No Knowledge Reply, we have:

1. The only precise credences I could form concerning the utility I assign to being a parent do not count as knowledge, because my statistical evidence doesn't allow me to rule out alternative reference classes that are made salient, or relevant, by the high-stakes decision I wish to make based on those credences;

2. By the Probabilistic Knowledge Norm for Reasons, these credences can therefore not provide a reason for me to act in any particular way, so that if I choose to do whatever maximizes expected utility relative to those credences, my reason for choosing in that way cannot be that the choice maximized expected utility for me, since that invokes my credences as a reason;

3. By the Probabilistic Knowledge Norm for Decisions, I am not necessarily required to choose the action that maximizes expected utility relative to those credences—they do not correspond to the strongest probabilistic content I know, and thus

[4] For more on the correct decision theory for imprecise credences, see Seidenfeld (2004); Seidenfeld etal. (2010); Elga (2010); Joyce (2010); Rinard (2015).

what is permissible for me is not determined by maximizing expected utility with respect to them.

What, then, am I required to do? That depends on what my statistical evidence allows me to know, and what the correct decision theory is for imprecise credences. As I mentioned already, there are many candidate theories, including the liberal and conservative versions described above. And on the question of what my statistical evidence allows me to know, we will have more to say below.

7 Assessing Moss's No Knowledge Reply: The Paulian View

We have now seen Paul's Utility Ignorance Objection to decision theory, the Fine-Graining Response, Paul's Authenticity Reply, and Moss's No Knowledge Reply. Given this, we can ask two questions: Does Moss's reply work from Paul's point of view? Does Moss's reply work independently of Paul's point of view? Paul emphasizes four important features of her objection. As we will see, Moss's reply to the Fine- Graining Response preserves two of those to some extent and two not at all. We begin with those it doesn't preserve.

First, Paul claims that the challenge to decision theory raised by ETEs is unique to those experiences. Whatever problem they raise, it is not raised by any other sort of phenomenon. And yet that isn't true on Moss's interpretation. Consider the doctor who must choose a treatment for her patient. She has the following statistical evidence: in 98% of trial cases, the treatment cures the illness; in 2% of trial cases, the patient deteriorates severely. She sets her credences in line with that. The illness is serious, so this is a high-stakes decision. Thus, other reference classes are relevant, and the doctor's evidence cannot rule out that the frequency of successful treatment is very different in those. So, by Moss's lights, the doctor's credence of 0.98 that the treatment will succeed and 0.2 that it will fail do not count as knowledge and so cannot provide a reason for action. Now, you might consider that the wrong conclusion or the right one—you might think, for instance, that the doctor's credences can provide reason for action, even if the doctor would prefer to have better evidence. But that is not the issue here. The issue is only that this other decision faces exactly the same problems that, for Moss, any decision faces that involves ETEs. That is, ETEs do not pose any new or distinctive problem for decision theory. And thus, on Moss's account, we lose this crucial feature of Paul's account.

The second distinctive feature of Paul's account is that, in decisions that involve ETEs, the problem is first-personal. When I am choosing whether or not to become a parent, the problem arises, according to Paul, because I am trying to make a decision for myself about my own future and yet I cannot access a part of my self that is crucial to the decision, namely, my utilities. This is why Paul turns to concepts like alienation

and authenticity to account for the phenomenon: they apply to first-personal choices in a way that they don't to third-personal ones. However, as the example of the doctor from above shows, there is nothing distinctively first-personal in Moss's diagnosis of the problem with decisions that involve ETEs—the problem arises just as acutely for a doctor making a major decision for a patient as it does for me when I try to choose whether or not to adopt.

The first feature of Paul's account that Moss's No Knowledge Reply does preserve and explain, though for quite different reasons, is the importance of what is at stake in the decision that we wish to use our credences to make. As Paul and Moss both acknowledge, there are trivial ETEs and important ones. When I choose whether to spread Vegemite or Marmite on my toast—having tried neither—I am choosing which ETE to have. But neither thinks that this poses a problem for decision-making in the way that choosing to become a parent does. Both think it is quite acceptable to use statistical evidence about the utilities that others assign to eating those two condiments as reasons I might cite when making my decision. Paul's explanation: only in significant life decisions do alienation and inauthenticity threaten. Moss's explanation: in low-stakes cases, there are no alternative reference classes that are relevant, and so my credences will constitute knowledge even if my evidence cannot rule out any alternative reference classes. Different explanations, but both agree that stakes matter. And both agree that, in these low stakes case, the fine-graining response is acceptable.

The second feature of Paul's account that Moss's reply preserves, though again for quite different reasons, is the attitude to decision theory. It is important to note that neither Paul nor Moss wish to abandon the machinery of decision theory in the face of the Utility Ignorance Objection; neither wishes to reject expected utility theory. Rather, in the case of significant life decisions that might give rise to ETEs, they advocate changing the decision problem that we feed into that decision theory. For instance, on the Fine-Graining Response, when I am deciding whether or not to adopt a child, I formulate the following decision problem:

the possible acts are

- Apply,
- Don't Apply, the possible states are
- Succeed & utility of being a parent is —12,
- Succeed & utility of being a parent is 3,
- Succeed & utility of being a parent is 10,
- Fail,

the doxastic states are my precise or imprecise credences over those states, on the supposition of those acts;

the conative states are my utilities over the conjunctions of acts and states, which encode the overall value I attach to these conjunctions, incorporating the quality of the phenomenal experience they give me, the moral and aesthetic values they boast, and so on.

I then feed this decision problem into the machinery of decision theory, which then tells me which of the possible acts are permitted by rationality and which are not.

For Paul, the new decision problem that we feed into the machinery of decision theory is this:

the possible acts are

- Apply,
- Don't Apply, as before;

the possible states are

- Succeed,
- Fail,

the doxastic states are my precise or imprecise credences over the states, on the supposition of the acts.

the conative states are my utilities over the conjunctions of acts and states, but instead of encoding the overall value I attach to these conjunctions, which Paul has shown we cannot access prior to making the decision, they encode only the value I assign to the revelatory experiences involved in those conjunctions.

This last element is the crucial ingredient in Paul's solution to the Utility Ignorance Objection (Paul 2014: 113)—when your utilities are not epistemically accessible to you, you should choose based on your preference for revelation; you should choose based on the utility you assign to discovering what it's like to have a new experience. I raised some worries about it in my review of Paul's book (Pettigrew 2016: 930-31).

Thus, the conative state specified in the decision problem is different from that in the orthodox version, while the doxastic state remains the same. In contrast, for Moss, the new decision problem is this:

the possible acts are

- Apply,
- Don't Apply, the possible states are
- Succeed & utility of being a parent is —12,
- Succeed & utility of being a parent is 3,
- Succeed & utility of being a parent is 10,

- Fail,

the doxastic states are not my precise or imprecise credences over the states, but rather the strongest imprecise states that count as knowledge for me, the conative states are my utilities over the conjunctions of acts and states, which encode the overall value I attach to these conjunctions, as in the orthodox approach.

Thus, the doxastic state specified in the decision problem is different from that in the orthodox version, while the conative state remains the same.

So, again, Paul and Moss agree—the orthodox decision problem should be replaced. But they agree for different reasons—Paul thinks that the conative state should be specified differently, while Moss thinks the doxastic state should be specified differently.

8. Assessing Moss's No Knowledge Reply: The Independent View

In this section, we continue to consider Moss's No Knowledge Reply to the Fine-Graining Response to Paul's Utility Ignorance Objection to orthodox decision theory. But this time we consider it independently of its relationship to Paul's own reply to that response to her objection. We can read Moss's No Knowledge Reply in one of two ways. On the one hand, granted the possibility of probabilistic knowledge and the accompanying probabilistic versions of the knowledge norms for action—Moss's Probabilistic Knowledge Norm for Reasons and Probabilistic Knowledge Norm for Decisions—we can read it as trying to establish that the Fine-Graining Response is wrong. On the other hand, if we start at the other end and assume that the Fine- Graining Response is wrong, then the need to appeal to probabilistic knowledge to explain why it is wrong is supposed to furnish us with an argument in favor of probabilistic knowledge, its possibility, and its use as a concept in epistemology.

The first worry I describe concerns the second reading. I will argue that a notion of probabilistic knowledge is not, in fact, required in order to explain the problem with decisions involving ETEs in the way Moss wishes to. The explanation can be given better, in fact, using only the familiar notion of probabilistic justification. The central point is this: the feature of first-personal utility credences based on statistical evidence that prevents them from counting as knowledge on Moss's account also prevents them from counting as justified.

In the Fine-Graining Response outlined in Section 4, I have credence 0.3 in —12, 0.6 in 3, and 0.1 in 10. I base these credences on my statistical evidence that 30% of parents assign utility —12 to being a parent, 60% assign utility 3, and 10% assign utility 10. Moss claims that these credences do not count as knowledge. I claim that, if they don't, they also don't count as justified.

Moss claims that these credences don't count as knowledge because my evidence doesn't allow me to rule out alternative reference classes that are rendered relevant by

the high stakes of the decision I am making. I claim that they don't count as justified for the same reason. After all, the ability to rule out relevant alternatives is important for justification too. Suppose Charlie and Craig are identical twins. I know this; I've known them for years. I also know that I can't tell them apart reliably. I see Craig in the supermarket and I form the belief that Craig is in front of me. Now, while true, my belief does not count as knowledge because I can't rule out the relevant alternative possibility that it is Charlie in front of me, not Craig. But equally my inability to rule out this possibility of which I'm fully aware also renders my belief unjustified. In general, if I believe p and there is an alternative possibility to p such that (i) I'm aware of it, (ii) I'm aware that it's relevant, and (iii) I can't rule it out, then my belief in p is not justified. The cases in which my inability to rule out an alternative precludes knowledge but not justification are those where either I am not aware of the possibility or not aware that it is relevant. For instance, in Goldman's Fake Barn County example, either I am not aware of the possibility of barn facades—perhaps I've never heard of such a thing—or, if I am aware of that possibility, I am not aware that it is relevant—because I don't know that I am in Fake Barn County. Thus, while I might be justified in believing that the structure in front of me is a barn, my belief doesn't count as knowledge. However, as soon as I learn about the possibility of barn facades and learn that I'm currently in Fake Barn County, my belief is neither justified nor knowledge. And the same goes for my credences about my utilities in the case of ETEs. Almost whatever statistical evidence I have about my utilities for becoming a parent, there is some relevant alternative reference class in which there are different frequencies for the various possible utility assignments such that (i) I'm aware of that reference class, (ii) I'm aware it's relevant, and (iii) I can't rule it out. Thus, any precise credence that I assign on the basis of that statistical evidence is not justified.

Thus, it seems to me that Moss's diagnosis of the problem with the Fine-Graining Response is wrong. The problem is not that the credences based on statistical evidence are not *knowledge,* it's that they're not *justified.* If that's right, then the argument in favor of the possibility of probabilistic knowledge that Moss bases on that diagnosis fails.[5]

But this seems a Pyrrhic victory. If I am right, surely this only makes the problem worse for the Fine-Graining Response itself. After all, the possibility of probabilistic knowledge and the putative norms that link it with reasons and decisions are contro-

[5] Of course, a knowledge-firster will claim that the credence based on the statistical evidence fails to be justified *because* it fails to be knowledge, not the other way around. While I am not a knowledge-firster myself, I think I can remain neutral on that claim here. I wish to say nothing about whether there is such a thing as probabilistic knowledge, nor if there is whether it plays the fundamental role in credal epistemology that the standard knowledge-firster claims non-probabilistic knowledge plays in the epistemology of non-probabilistic belief. I only claim that Moss cannot mount a certain sort of argument in favor of probabilistic knowledge, namely, that it is an essential ingredient in a plausible explanation of the difficulty of decision-making in the presence of epistemically transformative experience. That role can be played just as well by the notion of justification.

versial, whereas the possibility of probabilistic justification and the norms that link it with reasons and decisions are not. I think most decision theorists would agree that, while there is a sense in which an agent with unjustified credences should maximize expected utility with respect to those credences, such an agent will nonetheless not be fully rational. Thus, we seem to be left with a stronger reply to the Fine-Graining Response than we had before: we might call it the *No Justification Reply*.

But this is too quick. All that the considerations so far have shown is that, if I take a single statistical fact based on the distribution of utilities amongst people in a single reference class, and set my credences about my own utilities exactly in line with that, without considering anything else, then those credences will typically neither be knowledge nor justified. But there are other, better ways to respond to statistical evidence, and these can give justified credal states that can then be used to make our ETE decisions.

For instance, suppose I have the statistical evidence from above: 10% of all parents assign 10 utiles to being a parent, 60% assign 3 utiles, and 30% assign -12. But I also realize that I have properties that I share with some but not all parents: I enjoy spending time with my nieces and nephew; and I am a moderately anxious person. Let's suppose I think that the latter is the only property I have that affects the utilities I assign to being a parent. That is, I think that the distribution of utilities in the reference class of people who enjoy being around children is much the same as the distribution of utilities in the reference class of all parents, but the distribution amongst the reference class of moderately anxious people is quite different from the distribution in the class of all parents. And let's suppose that this belief is justified by my background evidence. Now, I don't know exactly what the latter distribution is, since that isn't included in my body of statistical evidence, but I have credences in the various possible distributions that are based on my background evidence. Let's assume again that those credences are also justified by my background evidence. I then use these credences, together with my statistical evidence concerning the distribution of utilities in the reference class of all parents, to set my credences concerning my own utilities for being a parent. The resulting credences will be justified.

Now notice: these credences will be justified not because I've ruled out the alternative distributions of utilities amongst the alternative reference classes, but rather because I've incorporated my uncertainty about those different distributions into my new credences concerning my utilities for parenting. And indeed that is the natural thing to do in the probabilistic setting. For many Bayesian epistemologists, nothing that is possible is ever completely ruled out; we just assign to it very low credence. This is the so-called Regularity Principle, and there are various versions determined by the various different notions of possibility (Shimony 1955; Stalnaker 1970; Lewis 1980; Jeffrey 1992). If the Regularity Principle is true, it is too demanding to require of an agent with probabilistic attitudes that they rule out alternative possibilities before they can know anything. Rather, we might say: in order for a probabilistic attitude to be justified, the agent must have considered all relevant alternative possibilities and

must have determined their attitude by incorporating their attitudes towards those possibilities. And we can do that in the case of credences concerning ETEs, even when those credences are based on statistical evidence, as we can see from the example of my adoption decision described above.

Now, I imagine that Moss might reply: while such credences might be justified, they will rarely count as knowledge. In order to count as knowledge, she might say, I must not only consider the properties I have that I think might affect the utility I assign to being a parent, and incorporate into my credences concerning that utility my uncertainty about the distribution of utilities for being parent amongst the reference classes defined by those properties; I must also consider the properties I have that will in fact affect that utility, and incorporate my uncertainty about the distribution of utilities for being a parent amongst the corresponding reference classes. Failing to consider those other properties might not preclude justification—I might be perfectly justified in not having considered those properties, and indeed justified in not even being aware of them. But it does preclude knowledge. Thus, just as I am perfectly justified in ignoring the possibility that the structure in front of me is a fake barn, but will be unable to know various propositions if that possibility is relevant in my situation, similarly, I might be justified in not considering various reference classes and the distribution of utilities within them, but nonetheless will be unable to know various probabilistic content if those reference classes are relevant in my situation. And thus, Moss might claim, by the Probabilistic Knowledge Norms for Action, the justified credences that I formed by incorporating my uncertainty about distributions amongst alternative reference classes cannot be used in rational decision making in the usual way.

The problem with this claim is that it asks too much of us. If, in order to know a probabilistic content concerning an event in a high stakes situation, you must have considered all of the causal factors that contribute to it being likely to a certain degree, there will be almost no probabilistic contents concerning complex physical phenomena that we'll know. In a high stakes situation, I'll never know that it's at least 50% likely to rain in the next ten minutes, even if it is at least 50% likely to rain in the next ten minutes, since I simply don't know all of the causal factors that contribute to that—and indeed, knowing those factors is beyond the capabilities of nearly everyone. There are many situations where, through no fault of our own, we just do not have the evidence that would be required to have credal states that count as knowledge. And this is not peculiar to credences concerning utilities for ETEs, nor even to credences based on statistical evidence.

Now, Moss might reply again: yes, it's difficult to obtain probabilistic knowledge; and perhaps we rarely do; and it's true that people shouldn't be held culpable if they violate the Probabilistic Norms of Actions; but that doesn't mean that we shouldn't strive to satisfy them, and it doesn't mean that the norms are not true. On this reply, Moss considers the Probabilistic Norms of Action as analogous to the so-called Truth Norm in epistemology, which says that we should believe only truths. Certainly, no

one thinks that those who believe falsehoods are always culpable. But nonetheless the Truth Norm specifies an ideal for which we should strive; it specifies the goal at which belief aims; and it gives us a way of assigning epistemic value to beliefs by measuring how far they fall short of achieving that ideal. Perhaps that is also the way to understand the Probabilistic Knowledge Norms for Action. They tell us the ideal towards which our actions should strive; and they give a way of measuring how well an action has been performed by measuring how far it falls short of the ideal.

But that can't be right. To see why, start by considering the following Non- Probabilistic Knowledge Norm for Reasons: a proposition p can count as your reason for performing an action just in case you know p. Now, that can legitimately be said to set an ideal, because there really is no extra feature of a categorical attitude towards p that we would want to add once we know p; it just doesn't get any better than that. The problem is that the same cannot be said in the case for probabilistic knowledge. Why? Well, suppose I know that it is at least 50% likely to rain. And suppose I am deciding whether or not to take my umbrella when I go outside. The higher the likelihood of rain, the higher the expected utility I assign to taking my umbrella. If it's over 40% likely to rain, I maximize my utility by taking it when I leave. Thus, since I know it's at least 50% likely to rain, I should take it. But this probabilistic belief concerning rain is not as good as it could be. If it's going to rain, it would be better if I were to believe that it is 100% likely to rain; if it's not going to rain, it would be better if I were to believe that it is 0% likely to rain. What's more, suppose I believe that it's at least 50% likely to rain. And suppose further that my belief is justified but not yet knowledge. Now suppose that I am going to gain one of two possible pieces of evidence. Either (i) I will gain evidence that turns my justified belief that it's at least 50% likely to rain into knowledge; or (ii) I will gain evidence that justifies a belief that it's at least 90% likely to rain, but will not turn that belief into knowledge. Which should I prefer, (i) or (ii)? If probabilistic knowledge is the aim of our probabilistic beliefs, and probabilistic knowledge is the ideal at which we should strive when form beliefs that ground our decisions, we should prefer (i)—we should prefer to obtain knowledge that it's at least 50% likely, rather than justified belief that it's at least 90% likely. But, I submit, (ii) seems just as good, if not better.

Before we wrap up, I'd like to draw attention to one final point, which is apt to be neglected. On the orthodox version of decision theory, an agent is bound to choose in line with her credences and her utilities—in the precise version of decision theory, for instance, she must pick an act that maximizes expected utility by the lights of her current precise credences. Both Moss and Paul argue that this is too demanding in the case of an agent who has adopted the Fine-Graining Response and who sets her credences in the fine-grained states in line with the statistical evidence. Requiring that she chooses in line with her credences, Paul argues, is tantamount to requiring that she makes her decision by deferring to the utilities of others—and that way inauthenticity and alienation lie. For Moss, on the other hand, it is not reasonable to demand that

an agent choose in line with beliefs in certain probabilistic contents—which is, after all, what her credences are—when she cannot rule out other probabilistic contents.

However, it is worth noting that the demand that orthodox decision theory makes is in fact rather weak. Suppose **P** is the set of credence functions that represents the strongest probabilistic content that you know. Then, in many cases, and certainly the cases under consideration here, **P** is also the set of all and only the credence functions that you are justified in adopting. Then, while it is true that, once you have picked your credence function P from **P**, you are bound to maximize expected utility with respect to P, you are not bound to pick any particular credence function from **P**—you might pick P, but equally you might pick any other P /' P from **P**, and you would be equally justified whichever you picked. Thus, the set of permissible choices for you is in fact exactly the same according to the orthodox view and according to Moss's Probabilistic Knowledge Norm for Decisions, when that is coupled with a liberal decision theory for imprecise credences. In each case, an act is permissible if there is a credence function P in **P** such that the act maximizes expected utility from the point of view of P.

I conclude, then, that Moss's No Knowledge Reply to the Fine-Graining Response does not work. I agree with Moss that credences based directly on sparse statistical evidence do not constitute probabilistic knowledge. But I argue that they are not justified either. And it is their lack of justification that precludes their use in decision-making, not their failure to count as knowledge. What's more, there are ways to set credences in the light of purely statistical evidence that gives rise to justified credences. Moss may say that these do not count as knowledge, and I'd be happy to accept that. But if she then also demands that credences used in decisionmaking should be knowledge, I think the standard is set too high. Or, if she thinks that probabilistic knowledge simply serves as an ideal towards which we ought to strive, then there are times when I ought to abandon that ideal—there are times when I ought to pass up getting closer to knowledge in one probabilistic content in order to get justification in a more precise and useful probabilistic content.

9. Conclusions

In the end, then, I conclude that the Fine-Graining Response to Paul's Utility Ignorance Objection to decision theory is safe. Paul's Authenticity Reply does not work, as I have argued elsewhere. And nor does Moss's No Knowledge Reply, as I have argued here.

References

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

Dougherty, T., S. Horowitz, and P. Sliwa. 2015. "Expecting the Unexpected." *Res Philosophica* 92(2): 301-21.
Elga, A. 2010. "Subjective Probabilities Should Be Sharp." *Philosophers' Imprint* 10(5): 1-11.
Goldman, A. 1976. "Discrimination and Perceptual Knowledge." *Journal of Philosophy* 73: 771-91.
Harman, E. 2015. "Transformative Experience and Reliance on Moral Testimony." *Res Philosophica* 92(2): 323-39.
Jackson, F. 1986. "What Mary Didn't Know." *Journal of Philosophy* 83(5): 291-5.
Jeffrey, R. C. 1983. *The Logic of Decision.* Chicago: University of Chicago Press.
Jeffrey, R. C. 1992. *Probability and the Art of Judgment.* Cambridge: Cambridge University Press.
Joyce, J. M. 1999. *The Foundations of Causal Decision Theory.* Cambridge: Cambridge University Press.
Joyce, J. M. 2010. "A Defense of Imprecise Credences in Inference and Decision Making." *Philosophical Perspectives* 24: 281-322.
Kahneman, D., and A. Tversky. 1979. "Prospect Theory: An Analysis of Decision Under Risk." *Econometrica* 47(2): 263.
Lewis, D. 1980. "A Subjectivist's Guide to Objective Chance." In R. C. Jeffrey (ed.), *Studies in Inductive Logic and Probability,* 263-94. Berkeley: University of California Press.
Moss, S. 2013. "Epistemology Formalized." *Philosophical Review* 122(1): 1-43.
Moss, S. 2018. *Probabilistic Knowledge.* Oxford: Oxford University Press.
Paul, L. A. 2014. *Transformative Experience.* Oxford: Oxford University Press.
Paul, L. A. 2015a. "Transformative Choices: Discussion and Replies." *Res Philosophica* 92(2): 473-545.
Paul, L. A. 2015b. "What You Can't Expect When You're Expecting." *Res Philosophica* 92(2): 149-70.
Pettigrew, R. 2015. "Transformative Experience and Decision Theory." *Philosophy and Phenomenological Research* 91(3): 766-74.
Pettigrew, R. 2016. Review of L. A. Paul's *Transformative Experience. Mind* 125(499): 927-35.
Quiggin, J. 1993. *Generalized Expected Utility Theory: The Rank-Dependent Model.* Dordrecht: Kluwer Academic.
Rinard, S. 2015. "A Decision Theory for Imprecise Probabilities." *Philosophers' Imprint* 15(7): 1-16.
Savage, L. J. 1954. *The Foundations of Statistics.* Hoboken, NJ: John Wiley.
Seidenfeld, T. 2004. "A Contrast Between Two Decision Rules for Use With (Convex) Sets of Probabilities: Gamma-maximin Versus E-admissibility." *Synthese* 140: 69-88.
Seidenfeld, T., M. J. Schervish, and J. B. Kadane. 2010. "Coherent Choice Functions under Uncertainty." *Synthese* 172: 157-76.

Shimony, A. 1955. "Coherence and the Axioms of Confirmation." *Journal of Symbolic Logic* 20: 1-28.
Stalnaker, R. C. 1970. "Probability and Conditionals." *Philosophy of Science* 37: 64-80.
Wakker, P. P. 2010. *Prospect Theory: For Risk and Ambiguity.* Cambridge: Cambridge University Press.

6. What Is It like to Have a Crappy Imagination?[9]

Nomy Arpaly

1. Introduction

The problem with human imagination is really two problems. One is that our imagination is very limited. Most importantly for this work, we have a very limited ability to imagine the lives of others and, by extension, future selves and potential selves. With others, it is the cause of endless misunderstandings, and the intrapersonal version of the problem causes a lot of surprises and bad choices. The second problem is that we trust our puny imagination a great deal, enough that we accept its testimony despite perfectly good evidence to the contrary. Thus, for example, an academic with a young child might say things like: "They told me that I won't get any research done the first few months but ... I guess I didn't take them seriously?" In other words, he received reliable information but dismissed it because his imagination told him he could just get work done when the baby is asleep.

My favorite cases involve disbelieving a person when she talks about the way she feels simply because one cannot imagine feeling as she claims to feel. Far be it from me to think that people are never wrong about their inner lives. Scientific studies can provide reasons to doubt, for example, that we remember our dreams as well as we think we do. However, when a person tells you she feels something, the simple fact that you can't imagine feeling that way should be regarded as a bad reason to doubt her, if it should be regarded as a reason to do so at all. I once told a relative stranger that though I grew up in a certain country, I do not feel identified with it. The man said, "That's highly unlikely." I am still a little angry at the person telling me what I feel, but I don't have the right to feel superior to him. After all, for decades, I believed that any person who claims to be "full after a salad" is either lying to me or lying to herself. In my defense, there exist some people who confess that they have kidded themselves on this particular topic, but I admit that I failed to believe anyone at all who claimed

[9] Nomy Arpaly, What Is It like to Have a Crappy Imagination? In: Becoming Someone New: Essays on Transformative Experience, Choice, and Change. Edited by: Enoch Lambert and John Schwenkler, Oxford University Press (2020). © Nomy Arpaly.
DOI: 10.1093/oso/9780198823735.003.0007

to feel satiated after eating only a salad simply because I can't imagine feeling that way—again, a bad reason.

This chapter is about the epistemic impact of experience. Experience has immense epistemic power—we will not talk here of its other powers, fascinating though they are—and there are many, many contexts in which we feel that there is no epistemic substitute to having "been there," as when you seek people who have suffered the same problems you have because nobody else truly understands, when you wish, say, that male politicians were forced to spend one day as women, or when you finally do experience something that you have prepared for endlessly and find yourself utterly surprised. This power that experience has can lead a philosopher to posit that there is a special kind of "knowing what it is like," different from "knowing that" or knowing facts, that only experience can provide. The question I would like to pose is how far one can go in accounting for the epistemic power of experience without resorting to this view. I will argue that to a large extent, in a large number of cases, the epistemic power of experience can be explained by the fact that it provides an antidote to the Problem with Human Imagination. That problem, in turn, is often enough to explain the seemingly insurmountable epistemic barrier we face when we try to figure out what a potential future self would be like, as in the cases, discussed by L. A. Paul (2014), of choosing whether to become a parent or whether to become a vampire.

One question I will not discuss here is the general question of how one is to decide what to do when one knows that at least one choice will result in a change to one's desires.[1] This question certainly arises in Paul's cases of becoming a parent or a vampire—parents and vampires have different desires from the rest of us—but I will assume that there is an answer to this question per se and only be interested in such cases if there is an extra problem: that the agent seems not to know what it is like to have the expected desire. Not all cases of desire change are cases of this sort. You might, for example, predict that if you'll accept a certain job, marry a certain person, or socialize with a certain group it will make you ambitious again. You have been ambitious before, so you can imagine quite well what being ambitious would be like; and so while philosophical questions are raised by this case, questions about what role the potential self 's preference should feature in your decision, these questions are not epistemic and not fundamentally about experience.

2. The Problem of Human Imagination

There are (at least) two common types of situation in which the Problem with Human Imagination interferes with our ability to understand the lives of others. One is relatively simple. Another is more complex and involves a phenomenon I call Runaway Simulation. Let us start from the first type.

[1] The problem appears, among other places, in Derek Parfit's case of the young nobleman. See Parfit (1984:327).

The Devil, they say, is in the details, and one might call cases of the first type I will discuss Devilish Details cases. I occasionally refer to them as "Giraffe Cases," because of a meme I saw on the internet which asks if it ever occurred to you that when a giraffe has coffee, the coffee is cold by the time it reaches its stomach. It answers that of course it hasn't occurred to you, because "you only think about yourself." Giraffes do not drink coffee, but the life of any human being who isn't you contains analogous facts that seem to make sense once you discover them but which would hardly ever occur to you on your own. Every person's life is shrouded in a cloud of little facts like these, and this accounts for a fair amount of the opacity of people to each other.

I have read that if you are a poor child in America, you can fail a course at school because your parents can't afford crayons, or, later, because you don't have enough gas to drive to the library. If one has never been in a similar situation one cannot simply imagine these facts. Now think of how many little facts like that there are— facts about being a poor child in America that, if you have always been rich or middle-class, or if you grew up in a country that has a much better safety net, haven't occurred to you any more than it has occurred to you that coffee would get colder before it reaches the stomach of a giraffe. While ignorance of each of these little facts feels like a simple case of failing to "know that," the cloud of ignorance that consists of not knowing many, many such facts can give a distinct impression that the child of affluent parents has about as much of an idea of what it is like to be a poor child in America as she has of what it is like to be a bat. Sadly, due to the nature of the Problem with Human Imagination, she is nonetheless fairly likely to think she knows what poverty is like, which can be a problem if she becomes, say, a lawmaker in charge of economic policy.

Devilish Details cases are relatively simple. More curious cases in which the Problem with Human Imagination raises its ugly head are cases of what I would like to call Runaway Simulation. When we try to understand another person, we often imagine ourselves in her situation. Many psychologists and philosophers have referred to this method as "simulation" and have taken it to be central to the way we "read" other people.[2] I will use the term "Runaway Simulation" to refer the all too common process in which the working assumption that, in a specific situation, the other person would do, think, feel, or want the same thing that you would do, think, feel, or want stops being a working assumption and becomes instead a stubborn belief that resists glaring counter-evidence. I am especially interested in cases where the working-assumption-turned-belief is that the observed person will not, in a particular situation, do or think or feel or want something that the observer, in the same sort of situation, never would. The case involving me refusing to believe that other people can sometimes feel full after eating only a salad, and believing so simply because I never feel that way, is a case of Runaway Simulation, and it can be used to demonstrate the way in which such

[2] I do not commit myself here to any of the theories espoused by these philosophers and psychologists, nor even to the general view that simulation is all, or most, there is to the way we "read minds." I am only committed to the commonsense view that we do, not too rarely, try to understand others by "putting ourselves in their shoes," whatever exactly that means.

stubborn false assumptions can snowball into even bigger misunderstandings. If you can't imagine anyone feeling full after a salad, you will find yourself believing very strange things about people who claim to feel that way. You will perhaps think they are liars, or that they are vain, or that they don't want to allow you to buy them more food—perhaps they can't wait to leave, or they hate owing favors? Alternately, you will suspect that they have an eating disorder, or at least some unrealistic ideal of freedom from lust and gluttony. Like a twisted Holmes, you will believe the improbable (she's lying) to avoid believing the impossible (she really is full after a salad, despite such fullness being unfamiliar to me). This is one way in which deep misunderstandings between humans can propagate quickly.

Sometimes Runaway Simulation is fairly harmless: you enjoy looking at pictures of your children, and so you fail to imagine that others might not enjoy looking at pictures of your children, or children in general, and so you post more pictures of your children in social media than your fellow users would prefer. At other times, I will shortly demonstrate, the misunderstandings that come from Runaway Simulation are nothing less than tragic. Before turning to that, though, one needs to distinguish Runaway Simulation—and, for that matter, the Problem with Human Imagination in general—from three things that can cause one to misrepresent other people in one's mind. These are: wishful thinking, sweeping generalization, and garden variety prejudices such as racism, sexism, and homophobia.

3. Three Things that Runaway Simulation Is Not

Start with wishful thinking. Wishful thinking, they say, is a type of motivated irrationality: The wishful thinker believes what she wants to be the case. Runaway simulation, despite the fact that it can lead to bad behavior, need not be motivated any more than the gambler's fallacy. It is, of course, possible for the person who misunderstands another due to Runaway Simulation to also be guilty of believing what she wants to believe. After all, we often want other people to be like us. The belief that other people will enjoy looking at pictures of your children is a natural result of Runaway Simulation, but it is also a belief that can come into existence through wishful thinking. However, the Problem with Human Imagination in general and Runaway Simulation in particular is a separate cause of false belief from wishful thinking (and also from elaborate self-deception, if such there be, in which one fools oneself into believing what one wishes to believe). This can be seen when one looks at cases where the Problem with Human Imagination leads agents to beliefs that are downright uncomfortable and disagreeable to them.

Let us look at a relatively low-stakes case. I am a night owl, for whom staying up until dawn is a common occurrence in the summer. Sometimes I talk to people and they start yawning. I feel bad: Am I boring them that much? Then, slowly, people start saying they have to go. I feel terrible: Did I say something wrong? Then someone

mentions the obvious fact that it's midnight and she is tired. I know, of course—"theoretically," as they say—that not everyone is a night owl like me, and yet, after all these years, my visceral feeling is something like "You're half my age. You can't possibly be tired at midnight. Are you sure I haven't said something wrong?"—and my credence that the lateness of the hour in fact explains my interlocutors' behavior is never as high as it should be. I detest the feeling that I am boring someone and detest even more feeling that I have done something wrong. Were I to believe what I wished to be true, I would have happily accepted the more flattering explanation for the yawns around me ("this person is tired") instead of the less flattering ("I am boring her to death" or "I said something wrong"). Yet, years since I have first noticed the phenomenon, I still cannot rid myself completely of my suspicion ("Good night! Wait, are you sure I didn't say something wrong?"). This is the work of Runaway Simulation.

Another cousin of the Problem with Human Imagination is good old sweeping generalization. A man who is indifferent to sports, a teenager who is not interested in teenage culture, any human being who prefers rainy days to sunny days or who hates sweets, all face the burden of having to incur, again and again, false perceptions and disbelief, simply because they are members of a small minority. But the Problem with Human Imagination is not the same problem as the human tendency to generalize. Unusual people have a hard time imagining the lives of mainstream people just as much as the other way around, though seeing representation of mainstream life on TV can be a partial (but very cheap!) substitute for imagining it. I often feel I have no idea what people in corporate offices do all day, as their working hours are full of Devilish Details to me—regardless of the fact that there are more corporate workers than academic philosophers. Runaway Simulation also goes in both directions. As I have mentioned earlier, I know very well that most people are tired at midnight, and yet when a person yawns at midnight in my presence I have trouble believing that he's not as awake as I am. A person I have known who never felt hunger was forever flabbergasted by the fact that her family wanted to eat three times a day ("Already? You're crazy!").

On to prejudice. Runaway Simulation cases and Devilish Details cases need to be distinguished from cases in which a person is opaque to another due to the latter's racism, sexism, or other prejudice of this type. The Problem with Human Imagination—whether we are talking about Devilish Details cases or Runaway Simulation cases—does not in itself discriminate against oppressed or marginalized groups. Just as a rich person would by default be likely to have a ridiculously false idea of what the life of the poor is like, due to a failure to imagine it well, a poor person will also by default have a ridiculously false idea of what the lives of the rich are like, for exactly the same reason. However, as the testimony of members of marginalized groups is widely regarded as less credible,[3] and as we often distrust people against whom we are prejudiced about as much as we trust our imagination, it is to be expected

[3] See Fricker (2007).

that a member of a marginalized group who is misunderstood by a more privileged person as a result of Devilish Details or Runaway Simulation will tend to have more trouble removing the misunderstanding—having her story believed, as it were—than a privileged person would have if he was himself so misunderstood.

4. Tragic Misunderstanding

Misunderstandings can be tragic, and I would like to take a more detailed look at the special way in which misunderstandings stemming from Runaway Simulation can ruin lives. For a few years, I volunteered in an informal setting to chat with people with mood disorders. I didn't keep statistics, but the most common, and most desperate, complaint the chatters had about their environment—even more common and more desperate than the complaints about the American health care system— was the complaint about people who fail to believe sufferers of depression that they in fact feel the terrible way they say they feel. Three common versions of the complaint are: "My husband is convinced I'm just doing it to get his attention," "My wife is convinced I'm just doing it to avoid my responsibilities," and "My parents think I should finally grow up and snap out of it." "It," in all of these statements, refers to acting like a person who is going through a moderate to severe episode of clinical depression.

The neurotypical party in these cases—usually a partner or a parent—often sticks to the conviction that the patient is faking (or, at best, semi-consciously, childishly dramatizing) his condition even though the patient has not been hiding feelings: the patient has been crying all day, getting into all manner of trouble for being unable to work or study, and avoiding favorite activities. He has confessed openly to feeling of deep despair. Yet all of this seems to go into one of the neurotypical party's ears and out the other. After all, she seems to think, nobody has died, the patient is not getting a divorce, his professional life is going fine, and so it doesn't stand to reason—that is, it is not imaginable to the neurotypical—that anyone in the patient's position is in fact as miserable as he seems to be. The disbelieving too often continues even after the patient protests that he would love to snap out of it, that he would if he could, that it hurts like hell. It persists after the patient specifies his feeling that he is an unbearable burden on his family and friends who would be happier and better off without him.

Sometimes, after the patient attempts suicide and goes through ghastly procedures such as stomach pumping, a diagnosis from the doctor in the hospital solves the matter. The doctor assures the neurotypical party that the patient is in fact suffering, and at times, our tendency to trust doctors trumps even our tendency to trust our imagination. At other times, even the "doctor's note" fails to help, and the partner or parent concludes from the suicide attempt that the patient would go very far indeed to get attention or avoid responsibility or avoid growing up—even risk death! "I wish you didn't waste all this time and money on doctors and medications," the partner or parent might say, "Just grow up already!" In short, in the very moments in which

sympathy is most urgently, desperately needed, the very moments in which lack of sympathy can be lethal, the at-risk patient is told that he is a manipulative bastard.

If one is a friend of such a patient and hears from her about the unsympathetic treatment she received, one is often inclined to be angry at the unsympathetic neurotypical party, and thus it is very tempting to find some moral fault with him. This, in turn, encourages us to think of the person's withholding of sympathy as somehow intentional, with seeing the patient as a manipulative bitch being a convenient excuse for him to act selfishly and "shut her out." Alternately, one might accuse the unsympathetic neurotypical of vice without attributing intentional motivation to her. "If she loved her more, she would have sympathized more," one thinks, or even "He would sympathize with him if he were capable of sympathy." At times one might think there is some kind of wishful thinking at play: the neurotypical party does not want to deal with the horrible reality of depression and prefers to think that the patient is manipulative or immature or both.

Despite the urge—which I personally feel—to shake the neurotypical who accuses the suffering depressive of manipulation or immaturity, Runaway Simulation ("I could never be so sad without a reason, so she can't be either") can explain many of these cases, and explain them in a way that is, as it were, charitable to the uncharitable.

Take the suggestion that the offending neurotypical refuses to believe the patient's suffering because of wishful thinking. In many cases, there are two basic problems with that explanation of the neurotypical's behavior. One has to do with fact that the content of what the neurotypical party does believe is not clearly more pleasant to her than the content of an accurate representation of the situation would be. Take the case of the partner whose belief is "she does it to get attention." The woman in question is believed to have faked, at a real risk to herself, a suicide attempt, complete with social stigma and medical costs, in order to make her partner fear for her, having already put on a long and elaborate charade intended to make people fear for her, on top of such things as refusing to do her share in the relationship or care for the children. If this were in fact the patient's behavior, it would indeed be manipulative and frightfully inconsiderate. If the neurotypical partner married her because he loved her, or at least liked her, is it that plausible to think that he now wishes to believe that the nice woman he chose to marry is now a terrible, shallow manipulator? The belief that one has married a manipulative bastard or bitch is not a comforting one, and it's not clear one would wish to think such a thing. The even more significant problem with the wishful thinking explanation is that the same person who refuses to believe that her partner is depressed without some kind of reason to be depressed, infuriating as she is in this case, can be perfectly sympathetic and display very caring behavior if her partner develops a problem (medical or otherwise) that she has no trouble imagining, and even more sympathetic if her partner develops a problem that she herself has experienced in the past. Given these two problems, and given the magnitude of the Problem with Human Imagination, it stands to reason that many of these tragic cases of misunderstanding can be explained by overconfident failure of imagination on behalf

of the neurotypical parties without resorting to the accusation of wishful thinking or convenient selfdeception.

Both Devilish Details and Runaway Simulation make it the case that one of the hardest things to imagine is having intrinsic desires (or likes and dislikes, or concerns) that you in fact do not have (especially, but not only, if you have the opposite desires). The lover of baseball ("How could anyone not love it?) and the person deeply indifferent to baseball ("It's like watching paint dry!") find it extremely hard to imagine each other, and an old joke states: "I am glad I hate spinach because if I didn't hate it I would eat it, and it's yucky!" I would like to point out that imagining having different intrinsic desires is even harder than it first seems.

I do not think that desire is itself an experience (though craving might be). However, having a desire greatly changes the way(s) one experiences the world. Most clearly, it influences what we experience as pleasant and what we experience as unpleasant—if you desire that a certain team win, you are happy when it does and sad when it does not. Elsewhere I argued that all of our experiences of pleasure and displeasure depend on our desires (even pleasure at eating a peach occurs only if you intrinsically desire certain taste experiences). A person with different desires will enjoy and suffer through different things, which is hard enough to imagine. This, however, is not all. Having or not having a particular, strong intrinsic desire matters to the cognitive world of the agent: it determines, to a large extent, what the agent notices, remembers, learns. To properly imagine the life of someone with a certain strong desire, one has to imagine the things that she would notice, remember and learn—and if one does not have the desire oneself, it's very, very difficult. Here is a somewhat detailed example. I love owls. I am not a knick-knack collector, but I desire to see owls (or at least their photographs) and to learn facts about owls. Due to this desire, I notice, for example, that the word "kn_owl_edge" contains the word "owl." Most people do not, even if they are philosophers and have seen the word "knowledge" in writing many times. A person who merely tries to imagine what it is like to be an owl lover would probably fail to see the "owl" in "knowledge." Having the desire, rather than just imagining having the desire, is usually required to notice such things. To the extent that she does not have the desire, a person who does not have the desire to see and learn about owls is unlikely to know how I see the world. Such opacity exists to an even greater extent when one has a considerably stronger desire to which a wider variety of circumstances is relevant, such as a parent's desire for the wellbeing of a child. To give an analogy to the "owl" example, it seems that very often, where I see a boring place, a parent sees a myriad of potential dangers to a child. Knowing that in principle does not, however, give me much of an ability to guess—or imagine—what potential dangers he sees. As the difference in desires between me and a loving parent is more significant in various ways than the difference between you and a person who loves owls, a parent's cognitive life, as well as her pleasures and displeasures, is even more opaque to me than mine are to you.

To recap: there is a Problem with Human Imagination, consisting of people having very limited imaginations but very high confidence in what their imaginations tell them. Devilish Details cases and Runaway Simulation cases are all cases in which the Problem makes it hard for us to understand other humans. But the Problem with Human Imagination does not only sow misunderstanding between existing humans. It also interferes with our attempts to understand future selves and potential future selves.

5. Imagining Who I Will Be

If one plans to become a parent, one might fail to predict a great mass of Devilish Details, and often does, resulting in shock and confusion. It might be less obvious how Runaway Simulation can occur when attempting to imagine a future self, but it is just as common. Consider the common warning against shopping while you are very hungry, as you will buy more than you need. In such a case, a hungry Monday self runaway-simulates a Thursday self and predicts that he be just as hungry, which results in his overestimating the amount of food that will be required on Thursday. Hilariously, a friend of mine suffers from the converse difficulty. Many times, when he packs lunch for the next day, he packs too little, because at the time of the packing he had just had dinner. If it is that easy to "over-simulate" your predictable next day self as resembling today's, it is all too easy to imagine your unfamiliar future or potential parental self as more similar to your current self than she has the right to be—especially as she has significant intrinsic desires that you now do not have. For example, when imagining a potential or future self who has a small child, it is still hard to imagine that self as able to read aloud from a book called Moo, Baa, La-La-La for the fortieth time in one weekend rather than as unable to stop screaming when a fifth time is called for.

The closest thing we have to a reliable cure for the ignorance imposed on us by the Problem with Human Imagination is experience. Normally, experiencing is believing. No matter how many years you have believed that nobody ever feels full after a salad, once you have felt so yourself—that is, experienced it—you will most likely think it is possible, after all. I say that experience is the closest thing we have to a reliable cure because it is not in fact that reliable a cure. People can fail to learn from experience because of tricks of memory ("I forgot just how painful I find these things") or just plain foolishness. To complicate things even more, there are conditions under which experience can change a person's beliefs but leave her visceral expectations intact. Mood swings provide a handy example. A person with mood swings—and these mood swings need not involve psychotic features—knows from experience that her moods will change, and yet depression, by nature, feels like it will last forever even if it is not severe enough to cause the person to believe that it will last forever. High moods, too, can make a person *feel* invincible—and so make it hard to imagine being in a low mood, even if one knows from experience what a low mood is like and that a low mood is likely

to descend on one at some point.[4] Despite cases like this, experience is fairly often the antidote to the ignorance created by the Problem with Human Imagination—which, as I pointed out at the beginning of this chapter, explains why people who have had a certain problem can feel so understood and "validated," as the colloquialism goes, in the presence of those who have had it too.

My proposal, then, is that the Problem with Human Imagination explains, in many cases, the power of experience to surprise us and our failures in anticipating what life would be like after a major change, especially one that involves a change in our desires. When I say "many cases," I mean to exclude primarily cases that involve exposure to radically different sensory stimuli, especially those involving acquiring a new sense modality, whether one is a deaf person acquiring the ability to hear or a person with typical sensory powers acquiring some of the abilities of a bat. Seeing color for the first time when one had seen no colors at all probably also falls into that category.

Why? Consider the rich person who is surprised to hear that a poor child in America can fail a course because she can't afford crayons. There is a sense in which "she should have seen" that: She knows that crayons cost money, that some school projects require crayons, and that schools in America do not give them away. The dots are there to be connected, and when the rich person discovers the fact, she might think something along the lines of "I suppose that would make sense." Or consider the man who says, "They told me I won't get any research done with a baby, but I guess I didn't take them seriously." This person is a bit comical. He *should have known* that "they" are probably right. Why should they be lying? In what relevant way is he different from them? Something similar is true about cases of Runaway

Simulation. The man who answered my statement that I don't identify with my place of birth with "It's highly unlikely" should have known better. Even the tragic cases can have this comical side. Imagine the person, the sort of whom is mentioned above, whose partner shows all the signs of severe depression and tells him about the depressed way she feels but who fails to believe that anyone can feel so bad if nothing is wrong. Sometimes, I have said, after a doctor diagnoses the partner with clinical depression and explains the concept to the couple, the neurotypical party would say, "I'm sorry, honey, I didn't realize what suffering you were going through." The partner would have a point if she replied, "But why didn't you realize it? I was telling you every single day!" Even where the majority of humans will suffer the same failure of imagination or a similar one, there is a sense in which the person whose imagination fails can be criticized. Things are different when a person is missing sensory experience. A blind person who cannot imagine seeing red or a typical person who cannot imagine using sonar like a bat cannot be accused of "not connecting the dots," as there are no "dots" available for her to "connect." She cannot be said to have a bad imagination any more than a person can be fairly said to lack physical fitness because she cannot fly like a bird.

[4] For an elaboration of the concept of visceral expectation, see Schroeder (2004: ch. 2).

When one decides to become a parent, or a vampire, one is clueless. Will some people be less clueless than others about the lives of other people and potential and future selves? The evidence suggests that the answer is that some people might be a little bit less clueless about some things than other people are. Some people have better imaginations: as an absent-minded and bumbling person, I can testify that Nabokov, who was neither absent-minded nor bumbling, imagined the inner life of the absent-minded bumbler Pnin remarkably well. Unlike Paul, I think that a truly outstanding book or movie can bring us closer to understanding other lives (though I agree that not every run-of-the-mill vampire book does!). Of course, even the most imaginative people can fail at imagining—in fact, not unlike the best baseball players, they fail at it most of the time. For example, it has been noticed by many recently that female characters in novels written by men are more likely than real women to enjoy things that straight men enjoy imagining—and one suspects, among other things, Runaway Simulation.

In addition to people who are better at imagining, some people—more of them—have less imaginative capacity than Nabokov but a better-than-average epistemic humility that occasionally rescues them from the pitfalls of trusting their imagination too much. When they are told that it's impossible for all but a few academics to both do their fair share of childcare and do research during the first year of their baby's life, they are forewarned—and when it happens, they are not surprised. Some people with depression are lucky enough to find one or two friends or relatives who never had depression themselves but who believe them when they report their feelings (and such a friend, who "gets" clinical depression but is not susceptible to it in the least, can be of immense help in many ways). When I asked many acquaintances whether they were surprised by the realities of parenthood, I received a wide range of answers. Some people reported a shock of sorts, a few reported no surprise at all, and most were in the middle, including one person who said: "I was surprised, but not, you know, Laurie Paul surprised or anything." One could stipulate that those who were unsurprised or less surprised were just lucky, but one must not rule out the possibility that some of the variety has to do with people having better or worse imagination and varied levels of epistemic humility.

By and large, though, we are clueless. Though I have argued that, philosophically speaking, there can often be a simpler explanation of Paul's phenomena than explanations that involve a special type of qualitative knowledge, I do not, practically speaking, have a lot of consolation for anyone who faces the choices Paul describes. We have a crappy imagination, we trust it too much, and we will forever be blown away or rudely awakened or otherwise blindsided by experience.

References

Fricker, M. 2007. *Epistemic Injustice: Power and the Ethics of Knowing.* Oxford: Oxford University Press
Parfit, D. 1984. *Reasons and Persons.* Oxford: Oxford University Press.
Paul, L. A. 2014. *Transformative Experience.* Oxford: Oxford University Press.
Schroeder, T. 2004. *The Three Faces of Desire*. Oxford: Oxford University Press.

7. What Imagination Teaches[(10)]

Amy Kind

1. Introduction

In "What Experience Teaches" (1988), David Lewis famously argued that "having an experience is the best way or perhaps the only way, of coming to know what that experience is like"; when an experience is of a sufficiently new sort, mere science lessons are not enough. Developing this Lewisian line, L. A. Paul has suggested that some experiences are epistemically transformative. Until an individual has such an experience, it remains epistemically inaccessible to her. No amount of book learning, or testimony from others, will be enough to give her this access—nor will be she be able to achieve it by way of imaginative projection. It's this last claim that I will question in this chapter. Can imagination teach us about fundamentally new kinds of experiences—experiences that are radically unlike any of the experiences that we've had before? As I argue here, this question should be answered in the affirmative.

2. Background: From Jackson to Lewis to Paul

Our story begins with Frank Jackson's paper, "Epiphenomenal Qualia," and more specifically, with one of the thought experiments contained therein (Jackson 1982). As part of his case against physicalism, Jackson asks us to consider Mary, a brilliant woman who has been confined to a black-and-white room for her entire life. Though Mary has normal color vision, she has never seen color. Mary lives at some point in the future in which we have achieved a completed color science, and while Mary is in the room, she learns the entirety of this science. Studying textbooks and viewing black-and-white lectures, Mary learns the complete physical story of color and color experiences—that is, she learns all of the physical facts. But now suppose that one day Mary is released from her black-and-white room and is able to experience color for the first time. Suppose, for example, that she is shown a ripe tomato. How will she react?

[(10)] Amy Kind, What Imagination Teaches In: Becoming Someone New: Essays on Transformative Experience, Choice, and Change. Edited by: Enoch Lambert and John Schwenkler, Oxford University Press (2020). © Amy Kind.
DOI: 10.1093/oso/9780198823735.003.0008

In particular, will she learn something? According to Jackson, it seems obvious that she will—that upon seeing the ripe tomato she'll have a sort of "Eureka!" moment—"Aha!" she'll say, "So that's what seeing red is like." But since Mary already knew all of the physical facts while in the room, whatever she learns cannot be a physical fact. Thus, the physical story about color and color experience leaves something out, and Jackson suggests that we can thereby conclude that physicalism is false.

In the 35 years since the paper was published, this argument—now commonly known as the knowledge argument—has been much discussed. Much of the discussion concerns whether and to what extent the argument succeeds in defeating physicalism. But this issue is not of concern to us here. Rather, for our purposes, what matters is the central intuition underlying the Mary case, namely, that absent color experience, Mary cannot know what seeing color is like. Interestingly, there has been widespread agreement about the truth of this claim. Though there are one or two notable exceptions, even the vast majority of knowledge argument opponents accept it.[1]

Consider, for example, the response to the knowledge argument offered by David Lewis. In "What Experience Teaches," Lewis argues that it would be a mistake to think that Mary gains any propositional knowledge when she leaves the room and sees the ripe tomato. Yes, she has an "Aha!" moment, and comes to know what seeing red is like, but this knowledge does not consist in the acquisition of any new fact; rather, what Mary gains is a cluster of abilities. In particular, Lewis claims that Mary gains the ability to imagine seeing red, to recognize seeing red, and to recall seeing red. This proposal, which has come to be known as the ability hypothesis, sees Mary's knowledge of color experience—and experiential knowledge more generally—not as propositional knowledge (knowledge that such and such) but rather as a kind of know-how.[2] Knowing what an experience is like simply consists of having these abilities. And, setting aside cases involving neurosurgery or magic, the only way to acquire this experiential knowledge—and hence to acquire these abilities—is to have the relevant experience itself.

Developing this Lewisian line, L. A. Paul has recently redeployed a version of the Mary case to motivate a paradox about what she calls transformative experience. For Paul, transformative experiences have both an epistemic and personal dimension. When an experience is epistemically transformative, you learn something that is in principle epistemically inaccessible to you absent that experience. When an experience is personally transformative, you undergo a change to your own point of view to such

[1] In a variety of works on the knowledge argument, both Paul Churchland and Daniel Dennett have repeatedly denied the intuition that Mary learns anything when she leaves the room, i.e., they argue that she can have knowledge of what seeing red is like even while she is inside the room. See e.g. Churchland (1985); Dennett (1991; 2007).

[2] The ability hypothesis was originally proposed by Laurence Nemirow (1980; 1990). Nemirow's version of this hypothesis focuses on only a single ability, namely, the ability to imagine seeing red. Lewis's version, which specifies the three abilities laid out in the text above, is now considered to be the standard version of the view.

a degree that even your core personal preferences may be changed. In such cases you may even take yourself to be so changed as to be essentially a new person. Moreover, as with cases of epistemically transformative experiences, the kinds of changes that result from personally transformative experience are in principle inaccessible to you prior to having the relevant experience.

When we consider Mary's release from the black-and-white room—and here Paul amends Jackson's original case so that we needn't suppose that Mary has any special scientific prowess or knowledge—it seems clear that her color experiences will be transformative in both the epistemic and personal sense. Following Paul, let's distinguish this case from Jackson's original case by referring to the protagonist as ordinary Mary. As Paul notes:

> Ordinary Mary does not and could not know what it is like to see color, and so she cannot know what it will be like for her to see color until she's left her room. In other words, ordinary Mary, before she leaves the room, is in a special kind of epistemic poverty, keyed to her inability to grasp crucial information of her future experiences. Once Mary leaves the room, her experience transforms her epistemic perspective, and by doing so, it transforms her point of view. (2014: 9-10)

While the case of ordinary Mary might seem removed from real life, it's nonetheless true that many of our major life experiences appear to resemble Mary's experience of color in being both epistemically and personally transformative. Consider having to grapple with the death of a parent, or of a child, or being the victim of a violent crime, or suffering a traumatic injury. Such experiences are often described by those who undergo them as "life-changing," a term that seems to capture their transformative nature. In all of the cases just listed, the relevant experience is simply thrust upon us, often without warning, but there are other transformative experiences that are presented to us as a matter of choice. For Paul, many of the major decisions that we confront as we go through life concern experiences that are both epistemically and personally transformative. When we face decisions about whether to become biological parents for the first time, or join the army, or undergo major surgery, Paul suggests that key facts about what the experience will be like are inaccessible to us prior to undergoing the experience itself; and she also suggests that the experience will profoundly change who we are, and do so in ways that cannot be predicted in advance. Thus we have the paradox that interests her: In cases involving transformative choice— that is, in cases where we're faced with deciding whether to undergo the experience— how is it possible to make such choices both rationally and authentically?

For our purposes here, just as we need not worry about the conclusions concerning physicalism that Jackson wanted to draw from the case of super-scientist Mary, we also need not worry about the conclusions concerning rational decision-making that Paul wants to draw from the case of ordinary Mary. Rather, what's important to us is

the claim underlying these cases, namely, that there are various cases of experiential knowledge in which the only way such knowledge can be gained is to have the relevant experience. In the remainder of this chapter, I will call into question the plausibility of this claim. My focus on what follows will be on epistemic transformativeness, though if my argument succeeds I suspect that much of it will be applicable to personal transformativeness as well. My argument will proceed, in effect, by turning the ability hypothesis inside out. Rather than thinking that we gain the ability to imagine a certain experience in virtue of knowing what such an experience is like, we should instead think that in at least some cases we can come to know what an experience is like in virtue of our imagining it.

3. Imagination and Decision-Making

Suppose that you're in the process of redecorating your living room. You've recently repainted the walls a rich shade of golden brown, and now you're at the furniture store trying to pick out a new sofa. Unfortunately, you forgot to bring paint chips or pictures, so you can't compare the sofa colors to the new color of the walls—and it would be a long trip home to retrieve them. So how do you make your decision about which sofa to buy, about which color sofa would look best in the newly repainted room? Here there seems an obvious solution: call upon your imagination. You imagine the various sofas in your living room, and you use these imaginative exercises to make the determination about which sofa would be the best choice.

This isn't the only context in which you use your imagination to help you in your decision-making. Suppose you're trying to decide whether to spend your Thanksgiving vacation in the desert of Joshua Tree or the mountains of Big Bear, or whether to get the iPhone Xs or the iPhone Xs Max. Or suppose, as you're headed out to spend an afternoon on the soccer sidelines watching your kids' games, you're trying to decide whether to bring the heavy EZ Up that offers more effective sun protection or the much lighter beach umbrella that's less effective at providing shade. In all of these cases, you are likely to be calling upon your imagination. Let's think about how you might go about your Thanksgiving vacation planning. You might first imagine hiking to the Lost Palms Oasis trail, returning home for a refreshing swim in the pool, and then having Thanksgiving dinner under the stars against a brilliant desert sunset. And you might next imagine skiing down the Widow-Maker slope at Bear Mountain, returning home for a relaxing soak in the jacuzzi, and then having Thanksgiving dinner inside in front a burning fire. And likewise for these other decisions: You might first imagine how much easier it will be to type and read on the bigger screen of the Xs Max, and then next imagine how much harder it will be to fit the phone in your pocket. Yes, you could go to the store for a hands-on test, but in the wee hours of the morning, when you're making the purchase online, there are no sample models available for you to try out.

In general, in making all sorts of decisions both big and small about our futures, we frequently call upon imaginative exercises to help us make better, more informed choices. Our imagination might not always get things exactly right, but it needn't be an infallible guide in order that we be rational in relying upon it. So one might thus naturally think that we could also call upon the imagination in the sorts of cases that Paul has in mind as epistemically transformative—cases where we are trying to decide whether to have a child, or to join the army, or to undergo major surgery. Just as our imaginative exercises give us epistemic access to the way the sofa will look in the newly repainted living room, or what's it like to spend Thanksgiving vacation in Joshua Tree, can't our imaginative exercises give us epistemic access to what it's like to be a parent?

For folks like Paul (and presumably Lewis) who answer "no," the difference between these ordinary decision-making contexts and the transformative decisionmaking contexts is that in the transformative cases we don't have enough experiential background to get an appropriate toehold for the imagination. Here we might recall a famous passage about the alleged inadequacy of imagination from Thomas Nagel's paper "What Is It Like To Be A Bat?" In the course of arguing that such knowledge is inaccessible to us, Nagel argues as follows:

> Our own experience provides the basic material for our imagination, whose range is therefore limited. It will not help to try to imagine that one has webbing on one's arms, which enables one to fly around at dusk and dawn catching insects in one's mouth; that one has very poor vision, and perceives the surrounding world by a system of reflected high-frequency sound signals; and that one spends the day hanging upside down by one's feet in an attic. In so far as I can imagine this (which is not very far), it tells me only what it would be like for me to behave as a bat behaves. But that is not the question. I want to know what it is like for a bat to be a bat. Yet if I try to imagine this, I am restricted to the resources of my own mind, and those resources are inadequate to the task. I cannot perform it either by imagining additions to my present experience, or by imagining segments gradually subtracted from it, or by imagining some combination of additions, subtractions, and modifications. (1974: 439)

Just as imagining that one has webbing on one's arms and that one has poor vision, etc., is not to imagine what it's like to be a bat, imagining that one has an infant strapped to one's chest in a baby carrier and that one is utterly exhausted is not to imagine what it's like to be a parent. And just as Nagel denies that the present experiential resources of someone who is not a bat are adequate to provide the appropriate materials for imagining what it's like to be a bat, Paul denies that the present experiential resources of someone who is not a parent are adequate to provide the appropriate materials for imagining what it's like to be a parent.

Importantly, these transformative cases are meant to be special ones. It won't be true in every ordinary case that one actually has to have undergone the experience

before to know what such an experience is like. If you haven't hiked the Lost Palms Oasis trail but you've hiked other trails in Joshua Tree, and you've seen an oasis before, or even just seen pictures of oases before, then by using your imagination—as Nagel says, by "imagining some combination of additions, subtractions, and modifications" to other experiences you've had—it's plausible to say that you could know what hiking the Lost Palms Oasis trail is like even without having done it. Let's call the process employed in this sort of case one of imaginative scaffolding. And this is the same kind of process that we use in the other kinds of decision-making cases mentioned above, from the living room redecoration to the iPhone purchase. Sure, I haven't ever held on iPhone Xs or Xs Max, but I've held an iPhone 7, and that seems to give me what I need to make the appropriate extrapolations.

Note that for our purposes here, we don't have to specify exactly what role this imaginative scaffolding plays in the production of this experiential knowledge, i.e. whether it is involved only in the context of discovery, or whether it can also be said to be involved in the context of justification. Historically, philosophers such as Jean- Paul Sartre and Ludwig Wittgenstein have famously argued that imagination never plays a justificatory role in our acquisition of knowledge. As Sartre claims, "nothing can be learned from an image that is not already known" (1948: 12). Since, on his view, "it is impossible to find in the image anything more than what was put into it," we can conclude that "the image teaches nothing" (pp. 146-7).[3] Wittgenstein makes a similar point when he notes that we are not surprised by the content of our imaginings (Zettel §632). Elsewhere I've argued against these claims; in my view, the fact that imagination calls upon our past experiences doesn't prevent it from being able to play a justificatory role in knowledge acquisition (Kind 2018). But in the present context, it doesn't really matter whether imagination gets the credit for providing justification of the relevant knowledge—all that matters is whether such knowledge is attainable. I'll often talk in terms of imagination providing the knowledge; this talk should be read as neutral on the question of the precise nature of the epistemic role that imagination plays in the process of knowledge-generation.

So there are some decision-making contexts—like redecorating the living room or vacation planning—where our experiential resources are sufficient to enable us to use imaginative scaffolding to gain experiential knowledge, whereas there are other decision-making contexts—like family or career planning—where they are not (or so it's been claimed). In the following section, I'll turn to the question of what exactly is supposed to account for this difference.

[3] See also McGinn (2004: 18).

4. Tasting Durian and Climbing Mountains

Let's consider another example of Paul's—that of tasting a durian for the first time. The taste of a durian is quite different from the taste of other types of fruit— indeed, it is quite different from anything most people have tasted before. Moreover, it is very hard to describe the taste of a durian to someone who hasn't tried one, and while some people love it, others find it completely repulsive. Indeed, the Internet is awash with videos of people trying durian for the first time, and their reactions can provide a few moments of diversionary amusement. Paul's thought is that before you've tasted durian, no matter how many different kinds of fruits or exotic foods you've tried in the past, you can't know what tasting durian is like.

Of course, when you're about to eat a strawberry, even when you've tasted strawberries many times before, there's also a sense in which you don't know what tasting this particular strawberry is like. Perhaps it will be slightly tart, or perhaps it will be especially sweet. Importantly, however, Paul doesn't want to treat tasting this particular strawberry as an epistemically transformative experience. As she puts it, "epistemically transformative experiences arise from having new kinds of experiences" (2014: 26).

Initially, this clarification may seem to help answer the question that we're here considering. Given that I've taken hikes before, the experience of hiking the Lost Palms Oasis isn't a new kind of experience, just like tasting this particular strawberry isn't a new kind of experience given that I've tasted strawberries before. But for ordinary Mary, who has never had any sort of color experience before, seeing red is a fundamentally new kind of experience. And likewise for the prospective soldier or the prospective parent. Given that a prospective soldier hasn't previously done anything like risking their life on the battlefield, and that a prospective parent hasn't previously done anything like bringing a new life into the world, such experiences also seem to be of fundamentally new kinds for the experiencer.

But matters are not quite this simple. Let's consider a case involving a considerably more elaborate hike than the 7.2 mile trek in Joshua Tree—namely, climbing Mount Everest. For many mountain climbers, climbing Mount Everest is a life-changing experience—one they claim can't be fully understood in advance. Take these remarks from climbing expert Alan Arnette, who summitted Everett on his fourth attempt at the age of 58: "Climbing Everest is hard. It tests you in ways you never knew possible.

You will understand that several months after you get home—regardless of your result If you summit, it will change your life. If you attempt it, it will change your life" ("I Want to Climb Mt. Everest," 2016). If we take Arnette at his word—and there's lots of testimony from other climbers who say similar things—then it looks like climbing Mount Everest should be a transformative experience.

But in what way is climbing Mount Everest a fundamentally new kind of experience? Consider an avid hiker and climber who has previously undertaken many mountain adventures. They have scaled Half Dome at Yosemite, they have spent three weeks hiking the John Muir trail, they have climbed Mount Whitney where the elevation at

the summit is 14,494 feet, they have climbed Mount Kilimanjaro where the summit is 19,341 feet, and so on. They have previously used ice axes and crampons and oxygen masks. So how is climbing Mount Everest a new kind of experience for them in the relevant sense? This experience falls under several kinds— mountain climbing experience, dangerous and strenuous mountain climbing experience, stretching oneself to the limit experience, and so on. But our seasoned climber has undergone all of these kinds of experiences before. Unless we have a very finegrained conception of experience—such that the relevant kind of experience is, simply, climbing Mount Everest experience—it's hard to specify an experiential kind that would be a fundamentally new one for our seasoned climber. And since climbing Mount Everest, i.e. the experience of climbing a particular mountain, sounds equally fine-grained as the experience of tasting a particular strawberry, it doesn't seem that Paul should classify this as a fundamentally new experiential kind.

Now Paul herself doesn't cite climbing Mount Everest as an example of transformative experience, and so presumably she'd simply deny that such an experience is epistemically transformative in her sense. But, as we'll see, the Mount Everest example is instructive, for it enables us to start to put pressure on what counts as undergoing a fundamentally new kind of experience. We have some cases that plausibly seem to be on one side of the line—like tasting a durian for the first time—and other cases that plausibly seem to be on the other side—like tasting a particular strawberry for the first time. In between we have a whole bunch of cases about which it's less clear what we should think. For someone who has had lots of varieties of melon—honeydew, cantaloupe, etc.—does it count as a fundamentally new kind of experience when she tries a canary melon for the first time? For someone who has had lots of apples and other types of fruit, does it count as a fundamentally new kind of experience when she first tries a pear? For someone who's tasted both apples and bananas, does it count as a fundamentally new kind of experience when she first tries a Hawaiian apple banana?

I won't here try to answer these questions, but I raise them to begin to put pressure on the idea that we can clearly demarcate some class of experiences that count as fundamentally new. This alone is not enough to show that Paul is mistaken, of course, since even if there are some cases that are hard to classify, as long as there is a class of cases that fall clearly on the side of being fundamentally new, Paul's arguments will hold for at least that class. In the next section, however, I will suggest that it's not even clear that there is such a robust class—or at least, it's not clear that the examples of transformative experience on which Paul focuses fall into that class. In what follows, I'll home in on one such example in particular, namely, that of having a child. Paul is explicit that this experience—which she understands as one of gestating, producing, and becoming attached to one's child—"is an experience with an epistemically unique character" (2014: 80-81). On her view, no matter how many new parents you talk to, no matter how much time you spend with your nieces and nephews, no matter how many parenting books you read, and no matter how much you try to imagine a parental future, you cannot know what it is like to become a parent prior to your becoming

so. As I will suggest, attention to cases of skilled imaginers helps to show why this is mistaken.

5. Imaginative Contortions

Having a child—gestating, producing, and becoming attached to the child—is an ongoing experience, one that has significant extension through time. In this respect it is unlike the experience of tasting a durian for the first time. So while we might think of the phenomenology of the durian experience in terms of some particular durian quale or qualia, it doesn't seem plausible to think of the phenomenology of the parenting experience in terms of some particular parent quale or qualia. Rather, the experience of having a child involves some complex and temporally extended phenomenology. When attempting to characterize the basis of this phenomenology, Paul notes that it could be the product of several different elements, either individually or in combination: the experience of producing a child, the experience of producing this particular child, your first experience of parental love, and so on. Because it is such an important part of her case, it here seems worth focusing a bit on parental love. I hope what I say will apply to the other aspects of parental phenomenology as well.

Assuming with Paul that you cannot experience parental love before becoming a parent, it's nonetheless true that there are lots of other varieties of love that you can experience. You might feel filial love for your own parents, romantic love for a partner, or so-called "brotherly" love towards other family members and friends. Some people love their close animal companions. Some have a love for humanity in general. And some people feel love for the god or gods that they worship.

Each of these varieties of love is different in various ways from the others—the love that you feel for your mother is different from the love that you feel for your best friend, and yet again different from the love that you feel for a romantic partner. But of course, the fact that the experience of hiking the Lost Palms Oasis trail is different from the experience of hiking the Bright Angel trail at the Grand Canyon does not mean that they're experiences of fundamentally different kinds. Someone who has done lots of hiking but who hasn't hiked the Bright Angel trail might be able to use imaginative scaffolding to discover what hiking that trail is like. Can someone who has experienced different varieties of love—who has loved her parents and her siblings and her friends and her Labrador retriever—but who has never experienced romantic love before use imaginative scaffolding to discover what romantic love is like? Certainly, little kids who watch princess movies and play dress-up and family take themselves to be imagining romantic love, as do hormonal teenagers who are yearning to find "the one." Or consider someone who has never shared her life with an animal before, and who is contemplating getting a canine companion. Is the experience of love that she'll

feel for her dog completely closed off to her, or might she be able to get some kind of good sense in advance?[4]

As different as these kinds of love are from one another, I'm disinclined to think that they are, in principle, imaginatively closed off from one another. Upon falling in (romantic) love for the first time, someone might plausibly say: "This is just what I've always been dreaming of." Granted, not everyone will say that. Some of the young children who imagine romantic love will not have gotten it right, and when they finally experience romantic love for the first time, they'll find that it's nothing at all like they'd imagined. Perhaps it's dramatically better, perhaps it's dramatically worse, perhaps it's more intense, or more all-consuming, or more selfless. But the fact that in some cases someone's experience of romantic love turns out to be different from how they'd imagined it to be does not show that the experience is imaginatively closed off, in principle, to anyone who hasn't experienced it. Rather, it might simply be that they're not very good imaginers.

Imagining is a skill. Like any skill, one can get better at it with practice, and some people are better at it than others. And also like any skill, what someone who has it can do can seem unfathomable to someone who lacks the relevant skill. Think of a skilled chef chopping vegetables, or a skilled juggler manipulating bowling balls and flaming clubs, or a skilled acrobat at a Cirque du Soleil performance. Once when I was in London on a family vacation, I saw a street performance by a performer who calls himself Yogi Laser. Yogi Laser is an amazing contortionist, and in addition to doing all sorts of poses with his body that I could hardly describe, in the finale of his act he folds himself up in such a way that he completely fits inside a glass box that measures 17 x 20 x 20 inches. This is a man who lists his height at 5'8 and his weight at 146 pounds. And apparently he holds the world record on this for speed—he's managed to get himself into the box in 5.35 seconds. (It took him a lot longer to do it in his street performance, but then again, he was also milking the crowd for donations the whole time.) Even seeing what he could do, I found it almost impossible to believe that someone could be so skilled.

Those of us who are not gifted contortionists might find it hard to fathom how a 5'8 person could fit himself into a 17 x 20 x 20-inch box. But the fact that we can't do it ourselves shouldn't lead us to conclude that it can't be done. Likewise, those of who are not gifted imaginers should be wary about drawing conclusions about what's imaginable from the fact that we can't successfully engage in particular imagining ourselves. The fact that some of us can't imagine a variety of love we have never felt before doesn't mean that no one can imagine a variety of love they have never felt before. Just as when we talk about what contortion is or is not possible for the human body we

[4] Here there is a big question about what exactly happens when someone who has never experienced romantic love learns via imagination what it is like, or when someone who has never experienced love for a canine companion learns via imagination what that is like. Is that exact same phenomenal quality produced? Many find this suggestion implausible. But what might happen instead? I regret that I don't here have answers to these questions, but I hope to return to the matter in future work.

should focus on skilled contortionists, when we talk about what imaginative exercise is or is not possible for a human experiencer we should focus on skilled imaginers.

Here it's worth noting that there are two different ways of being a skilled imaginer.[5] The first involves being skilled at fantastical imaginings. Someone with this skill, which relates to what I have elsewhere referred to as the transcendent use of imagination (Kind and Kung 2016), is very good at letting their imaginings run wild. The second involves being skilled at realistic imaginings. Someone with this skill, which relates to what I have elsewhere referred to as the instructive use of imagination, is very good at constraining their imaginings to fit the facts of the world. What's of interest to us here is this second type of imaginative skill.

Some imaginers who are particularly skilled in this second sense can do various design tasks in their imagination that others of us cannot do even equipped with computers, drawing tools, and physical models. Consider the inventor Nikola Tesla, who could design complex circuitry in his mind without ever putting pen to paper. As Tesla describes his design process:

> Before I put a sketch on paper, the whole idea is worked out mentally. In my mind, I change the construction, make improvements, and even operate the device. Without ever having drawn a sketch, I can give the measurement of all parts to workmen, and when completed these parts will fit, just as certainly as though I had made accurate drawings. (Tesla 1921)

Or consider the animal scientist Temple Grandin, who developed more humane forms of animal-handling equipment by taking what she called a "cow's eye view" of the situation and then running imaginative simulations in her mind. As she puts it:

> Now, in my work, before I attempt any construction, I test-run the equipment in my imagination. I visualize my designs being used in every possible situation, with different sizes and breeds of cattle and in different weather conditions. Doing this enables me to correct mistakes prior to construction. (Grandin 1995: 20-21)

Grandin's designs were so innovatively different from the designs previously in use that animal handlers were deeply skeptical that they would work, even when they saw the detailed plans for the equipment all drawn out. But such designs, once properly built, worked exactly as Grandin had predicted.[6]

Or, for just one more example, consider the master origami folder Satoshi Kamiya. In 2006, Kamiya produced "what is considered the pinnacle of the field, an eight-inch tall Eastern dragon with eyes, teeth, a curly tongue, sinuous whiskers, a barbed tail, and a thousand overlapping scales" (Kahn 2006: 60). Just the folding process itself

[5] Thanks to Lloyd Humberstone for encouraging me to draw this distinction.

[6] I discuss the examples of Grandin and Tesla more fully in Kind (2018).

took over 40 hours. Unlike other origamists, however, Kamiya produces his creations without the help of any software or computer aid. When asked how he can achieve such elaborate design without digital assistance, a feat that seems almost incomprehensible to his competitors, Kamiya's answer indicates the importance of imagination: "I see it finished. And then ... I unfold it. In my mind. One piece at a time" (Kahn 2006: 63).[7]

Above I was focusing on the possibility of knowing what romantic love is like even if one has never experienced it, and I think a plausible case can be made that someone can achieve such knowledge by way of imaginative scaffolding. The fact that not everyone can do it doesn't mean that no one can. In my view, we have no reason to think that things should be any different with respect to parental love.

Let's pause for a moment to consider how much people actually try to imagine parental love, prior to becoming parents themselves. How much do they work at it? In my own case, I have to confess, the answer was: Not very much. Imagining parental love was not a project that I had set myself. Presumably there are some folks who set themselves this task, but I suspect many others had similar experiences to mine.[8]1 suspect that what often happens to prospective parents is something like this: Perhaps when visiting friends who have recently become parents they see evidence of the bond those friends have with their newborn, and they spend a few moments marveling at the strength of it, wondering what it might be like to feel that way for someone else. And perhaps they even try to imagine it, only to conclude minutes (or perhaps only seconds) later that it's imaginatively out of reach.

But why should we think that a few minutes (let alone a few seconds) of effort are enough? Were we to draw this sort of conclusion with respect to other skills it would simply be laughable. And we do laugh, when, after just two or three tries at something a stubborn five-year-old stamps their foot and emphatically declares, "I can't do it." The amount of effort needed to acquire a new skill is often quite significant—and this can be so even when you have closely related skills. A few years ago, our family acquired a ping-pong table, and my older son—who was already quite good at putting spin on his shots—decided he wanted to try to master a trick serve that he'd seen in a YouTube video. Eventually he just about got it—he can't do it correctly every single time, but he can usually make it work in about one of three tries. To get to this point, however, he tried out that serve hundreds and hundreds (maybe even thousands and thousands) of times. Whether it's learning to juggle a soccer ball, or to play an intricate piano piece, or to perform some sleight of hand, acquiring a skill takes practice. And this practice often takes many forms—not only engaging in multiple attempts, but also breaking the task down into smaller parts, trying slightly different related tasks, and so on.

As we noted, being a parent is a complex experience. It's built of many parts and has a complex phenomenology. All together, these parts seem hard to reach. And even

[7] I first learned of Kamiya from the discussion in Grandin and Panek (2014).

[8] When I've presented this paper to various audiences, there are occasionally members of the

when we try to separate the parts, they still may seem hard to reach. But just as we acquire other skills by breaking them down into smaller parts, here we need to do the same thing, latching onto the little bits that we can, here and there, and then scaffolding our way onwards and upwards from there.

Indeed, this might make us rethink what we earlier said about tasting a durian. Above I said that it could plausibly be seen as a fundamentally new kind of experience. But even tasting a durian is a complex experience of sorts, and so perhaps it too can be broken down into smaller parts, each of which might be individually reachable by way of imagination, working—as Nagel said—to imaginatively add to, subtract from, and modify experiences that we've already had. One of the things that makes skilled imaginers better than unskilled imaginers is their ability to make these imaginative additions, subtractions, and modifications. The discussion in the previous part of the chapter suggested that it's a lot harder than it might initially seem to cordon off a class of experiences that are fundamentally or radically new in the sense that Paul is interested in. I've now suggested that even experiences that seemed clearly to fall into such a class may indeed be imaginatively accessible, at least by skilled imaginers. Once we attend more carefully to what can and cannot be done with imagination, it turns out that considerably fewer experiences remain imaginatively out of reach than proponents of transformative experience would have us believe.

6. Back to Mary

But here we might think that one kind of example remains—and indeed, it's the one with which we began, that of ordinary Mary. Mary's experience of seeing red for the first time seems unreachable in a certain way that these other kinds of experiences don't. And perhaps something similar could be said for experiences in new sensory modalities—as when a deaf person hears for the first time by way of a cochlear implant, or when (in a near or distant future) new technologies enable us to acquire electric or magnetic senses. Even if we can imagine parental love, in other words, it seems unlikely that we can imagine having an electro-magnetic sense. Though I feel the pull of this assessment, in this closing section, I want to briefly see whether we can put pressure on it, i.e. explore whether even this conclusion might be unfair to imagination. I don't have the space here to develop this argument at any length (though see Kind 2019), but let me make at least a start on trying to see whether and to what extent we can erode the difference between the Mary case and the other kinds of cases that we've been considering.

Here it will be helpful to consider an example that Paul Churchland offers in his own discussion of the Mary case, namely that of sightreading music. Many skilled musicians can sightread scores they've never heard before—i.e. even without hearing such scores,

audience who claim that they did seriously set themselves this task. Interestingly, however, in each such case, the person claimed that their advance imaginings got things pretty much right—in other words,

these trained musicians know what they would sound like when played. Extending this kind of example, Churchland notes that musicians with sufficient training can identify the individual notes of a chord they're hearing for the first time, and conversely, can auditorily imagine an unfamiliar chord from the specification of the notes. Such imaginative feats are possible in virtue of the fact that chords are structured sets of elements, i.e. in virtue of the fact that even new and unfamiliar musical experiences of chords are not that distant.

So now why should we think that color experiences are any different? The reason seems to be that color experiences, unlike chord experiences, seem to us to be undifferentiated wholes. But, as Churchland notes, many untrained listeners typically experience chords as undifferentiated wholes on first hearing them. So, asks Churchland, "Why should it be unthinkable that sensations of color possess a comparable internal structure, unnoticed so far, but awaiting our determined and informed introspection?" (1985: 26-7). If color sensations are likewise structured, then new and unfamiliar color experiences would be considerably less inaccessible to us, i.e. they'd be much easier to reach by way of imaginative scaffolding. The reason that we take color experiences to be imaginatively inaccessible to Mary is that there seems to be no way to get to them from the kinds of experiences she has. If color experience is structured, however—in particular, if it has some kind of phenomenal structure—then that structure could provide a way for Mary to scaffold out from the experiences she has to those that she hasn't. In doing so, this imaginative scaffolding could, in principle, help to teach her what seeing color is like.

There are a lot of "ifs" here, and so I don't take this discussion to have shown that Mary can imagine these things. Maybe she can't, even in principle. But importantly, even if knowledge of color experience is epistemically out of reach to her while she's in her black-and-white room, it will still be true that the Mary case remains an exception. More importantly, for the purposes of Paul's overarching project—one about the rationality of transformative choice—it's not a particularly interesting exception. What makes Paul's arguments in *Transformative Experience* (2014) so threatening is that they seem to show that we are unable to make rational choices in a vast array of the major decisions we face in everyday life. If, however, the problem is limited to cases that are analogous to the Mary case, much of the threat is dissipated. By focusing on the case of parental love, I have tried to show that many of the examples on which Paul relies are not analogous to the Mary case, i.e. they are not plausible cases of epistemically transformative experiences in the way that Paul thinks. Perhaps there are some epistemically transformative experiences—some experiences that are out of reach to us in advance, even imaginatively—but such experiences would seem to be considerably less common that we were led to believe, and they do not seem to be the kinds of experiences that matter for most of our everyday decision-making. Ultimately,

---

the experience of parental love turned out to be pretty much what they had imagined prior to becoming parents.

then, though it may be right that experience is the best teacher, as Lewis said in the passage with which we began, I hope here to have shown how, in a vast array of the cases that Paul considers, imagination can come in a close second.[9]

References

Churchland, P. 1985. "Reduction, Qualia, and the Direct Introspection of Brain States." *Journal of Philosophy* 82: 8-28.

Dennett, D. 1991. *Consciousness Explained.* Boston, MA: Little, Brown.

Dennett, D. 2007. "What RoboMary Knows." In T. Alter and S. Walter (eds), *Phenomenal*

Concepts and Phenomenal Knowledge, 15-31. Oxford: Oxford University Press.

Grandin, T. 1995. *Thinking in Pictures.* New York: Random House.

Grandin, T., and R. Panek. 2014. *The Autistic Brain.* New York: Mariner Books.

I Want to Climb Mt. Everest. 2016 (December 16). <http://www.alanarnette.com/ blog/2o16/ i2/i6/i-want-to-climb-mt-everest/>

Jackson, F. 1982. "Epiphenomenal Qualia." *Philosophical Quarterly* 34: 147-52.

Kahn, J. 2006. "The Extreme Sport of Origami." *Discover* 27(7): 60-63.

Kind, A. 2018. "How Imagination Gives Rise to Knowledge." In F. Dorsch and F. MacPherson (eds), *Perceptual Memory and Perceptual Imagination,* 227-46. Oxford: Oxford University Press.

Kind, A. 2019. "Mary's Powers of Imagination." In S. Coleman (ed.), *The Knowledge Argument,* 161-79. Cambridge: Cambridge University Press.

Kind, A., and P. Kung. 2016. "Introduction: The Puzzle of Imaginative Use." In A. Kind and

P. Kung (eds), *Knowledge Through Imagination,* 1-37. Oxford: Oxford University Press.

Lewis, D. 1988. "What Experience Teaches." *Proceedings of the Russellian Society* 13: 29-57.

Nagel, T. 1974. "What Is It Like To Be a Bat?" *Philosophical Review* 83: 435-50.

Nemirow, L. 1980. "Review of *Mortal Questions*, by Thomas Nagel." *Philosophical Review* 89: 473-7.

Nemirow, L. 1990. "Physicalism and the Cognitive Role of Acquaintance." In W. Lycan (ed.), *Mind and Cognition: A Reader*, 490-99. Oxford: Blackwell.

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

Tesla, N. 1921. "Making Your Imagination Work for You." *American Magazine* (April).

[9] For helpful discussion and feedback on previous versions of this paper, I am grateful to audiences at the 2017 annual meeting of the Southern Society for Philosophy and Psychology, the Pre-Conference on Transformative Experience at the 2017 Pacific APA, and the Imagination and Knowledge conference at the University of Konstanz. I am also grateful to the audiences at departmental colloquia held at

Wittgenstein, L. 1967. *Zettel,* ed. and trans. G. E. M. Anscombe and G. H. Wright. Oxford: Blackwell.

the Australian National University, Auburn University, the Institut Jean Nicod, and Monash University. For comments on previous drafts, special thanks go to Mark Steen, my commentator at the APA Pre-Conference, and to John Schwenkler and Enoch Lambert.

8. Transformative Activities[(11)]

Agnes Callard

> Would she always do the things I was supposed to do, before and better than me? She eluded me when I followed her and meanwhile stayed close on my heels in order to pass me by ... Maybe I should erase Lila from myself like a drawing from the blackboard, I thought, for, I think, the first time. I felt fragile, exposed, I couldn't spend my time following her or discovering that she was following me, either way feeling diminished. I immediately went to find her. I let her teach me how to do the quadrille.
>
> (Ferrante 2012: 142)

1. Introduction

In this passage of Elena Ferrante's My Brilliant Friend, the protagonist, Elena, describes her frustrations with the fact that she feels consistently bested by her friend Lila. This moves her to consider severing ties with Lila—but instead Elena decides, once again, to scramble to catch up with Lila. Despite the fact that Elena is the one who receives a higher education, she feels, "I knew little or nothing. She seemed ahead of me in everything, as if she were going to a secret school. I noticed also a tension in her, the desire to prove that she was equal to whatever I was studying" (2012: 160). Lila, too, is scrambling to catch up with Elena.

My Brilliant Friend is the first of four novels describing the lifelong friendship between Lila and Elena. In the various pursuits that become central to their lives over the course of the novels—reading and writing fiction, learning ancient (Greek and Latin) and formal (computer) languages, designing shoes, fomenting political unrest, romance, and child-raising—each is impelled forward by the thought that she is falling behind the other. These pursuits are transformative, in that the women gain an acquaintance with what it is like to do something of a different kind from anything they have done before, something that changes their preferences, attachments, and values. And these new lives are not foisted on Lila and Elena by their environment. Far from

[(11)] Agnes Callard, Transformative Activities In: Becoming Someone New: Essays on Transformative Experience, Choice, and Change. Edited by: Enoch Lambert and John Schwenkler, Oxford University Press (2020). © Agnes Callard.
DOI: 10.1093/oso/9780198823735.003.0009

it: At each point, Lila and Elena must fight the tide of cultural, familial, and economic pressure persistently pushing them in the reverse direction. Elena and Lila not only undergo transformative experience, but are also agentially responsible for those transformations.

And yet it is hard to find, in the roughly 2,000 pages Ferrante devotes to describing these transformations, the critical choice points that figure so centrally in the philosophical literature on transformative experience. The choices that Lila and Elena make in relation to the people they become do not qualify as what L. A. Paul has called "transformative choices." Look back at the passage I quoted in my opening, in which the decision to undergo a transformative change—to be taught to dance—is presented as an afterthought. Elena is not asking herself whether she wants to have the experiences and preferences of someone who knows how to dance. Her attention is, as always, on Lila. And this is typical of the various transformations that the women experience: Though they go on to lead radically different lives, each always has the other in view. I propose to examine their relationship as a case study in a distinctive—and distinctively competitive—form of transformative experience.

I begin with a taxonomy of transformative experiences, dividing them into what I call "revelations" and "activities." While most of the examples of transformative experience in the philosophical literature are examples of transformative revelations, Lila and Elena's transformations are of the other variety. Transformative activities present a different set of puzzles from those already explored in the literature focusing on transformative revelations. They require an agent to act in such a way that she may learn both what she is doing and why. Competition can facilitate this form of learning, allowing us to reach beyond our present set of resources, towards a future persona we do not have fully in view.

This tension is captured in the push–pull of Elena's relationship with Lila: "I felt fragile, exposed, I couldn't spend my time following her or discovering that she was following me, either way feeling diminished. I immediately went to find her. I let her teach me" (quoted above). Ifwe approach the story of Lila and Elena through the lens of the concept of transformative activity, we will find that descriptions such as this one afford us access to the inside of transformative experience—which is say, to what it feels like to be in the process of transforming oneself into a different kind of person.

2. Transformative Activity

In a transformative revelation, an agent elects to undergo some process which promises to provide her with new phenomenal knowledge, and perhaps also new preferences, desires, and values. Edna Ullmann-Margalit calls such a choice, "opting": "The opting juncture is a point of discontinuity, or break, in the opters' biography and personality" (2006: 168). The decision to become a new kind of person constitutes a dividing line between the person you were (cognitively or conatively) and the person you

will be. As an illustration, consider Paul's (2014) guiding example of the decision to become a vampire. The person in question elects to be bitten, and the bite transforms her into someone with a host of new experiences and desires: the desire for blood, an aversion to sunlight, excellent fashion sense, etc. Paul and Ullmann-Margalit approach the topic of transformative experience through the question of what it would be to make such a momentous choice in a rational way. For this reason, the literature on transformative experience has focused on examples of transformative revelations. But not all transformative experiences are transformative revelations.

In what I will call a transformative activity, the agent actively works to become a new kind of person without undergoing a break in her biography. So, for instance, learning a foreign language can be a transformative experience, affording a person access to a different way of thinking and conducting one's life. This sort of case seems to lack the "point of sharp discontinuity" that Edna Ullman-Margalit takes to characterize the transition from the "old person" who enters the transformative revelation and the "new person" who exits it (2006: 159). The student does not change much in the first month or even year of French class. And how different she becomes depends on how far she is willing to take it. Does she: take second and third and fourth year French? start reading French novels in French? major in French? visit France? move to France? attempt to assimilate to French culture? Such a person may, at the end of the day, have acquired new knowledge and preferences—the knowledge of what it is like to be French, and preferences characteristic of French people. She has undergone a transformative experience unlike that of becoming a vampire.

Let me illustrate my distinction with a pair of cases. The first I take from Paul: deciding to taste durian fruit.

> The durian is known to have a foul smell but a delicious flavor. I've never tasted a durian, and because I've never tasted one or anything like one, I can't know what it is like to taste one—that is, I can't know what a durian taste like When I taste it for the first time, by becoming acquainted with this taste, I'll undergo an epistemically transformative experiences, and gain new knowledge, the knowledge of what it's like to taste durian. (2014: 15)

Paul's attention here is on the fact that one cannot know what durian fruit is like until one has eaten it. But it is also true that one cannot but know what it is like once one has placed it into one's mouth. The knowledge of what the fruit is like is, as it were, inflicted by the fruit on our taste buds.

Contrast the tasting of durian with becoming a wine connoisseur. Tasting the note of cherry in some wine is an transformative experience: Before one has had this experience, one does not know what it is like for wine to have notes of cherry. Nonetheless, this experience is not something that simply happens to (most) people. I have drunk many glasses of wine, and doubtless some of them had notes of cherry— but I have never tasted it. The would-be connoisseur works to become such as to appreciate the subtle flavors and distinctions in wine—she takes classes, discusses wines, attends tastings, and, perhaps most importantly, she consciously attends to and tries to analyze her

own experience each time she has a glass of wine. The durian eater's transformative experience takes the form of a revelation, whereas the wine connoisseur's takes the form of an activity.

Not all examples are as clear-cut as the ones described above. For example: into which category should we place becoming a parent? Paul's description suggests that she takes the transformation in question to be a revelation:

> The intensity and uniqueness of the extended act of carrying the child, the physicality of giving birth... results in a dramatic change in one's physical, emotional and mental states. The experiences are also very intense for involved fathers. It is common for fathers to date their changed phenomenal state from the moment they saw or held their newborn. (2015: 156)

If, by contrast, we see parents as having a hand in both whether and how they grow into the role of parent—working to change their preferences and tastes in the direction of those befitting a parent, potentially culpable if they never manage to grow into the role—then we will see becoming a parent as a transformative activity.

Likewise, consider the example of becoming blind. Paul (2014: 70) gives blindness as one of those conditions that would be transformative for a sighted person: sighted people do not know what it is like to be blind. Her discussion as a whole treats the experience of losing or acquiring a sense modality (or aspects of a sense modality, such as color vision) as revelations. But consider this description from the diary of John Hull, a theologian and disability activist who recorded his own experience of going blind in his forties:

> As one goes deeper and deeper into blindness, the things which once were taken for granted, and which were then mourned over as they disappeared, and for which one tried in various ways to find compensation, in the end cease to matter. Somehow, it no longer seems important what people look like, or what cities look like. One cannot check at first hand the accuracy of these reports, they lose personal meaning and are relegated to the edge of awareness. They become irrelevant in the conduct of one's life. One begins to live by other interests, other values. One begins to take up residence in another world. I think that I may be beginning to understand what blindness is like. (1990: 192)

This entry is from October 1985, by which time Hull had been blind for two years. And yet he is only "beginning to understand what blindness is like." Hull's experience of coping with blindness is that of working his way into what he called "deep blindness"—abandoning mental images and even spatial concepts such as "here" and "there" so as to learn to think in a new way. Hull's understanding of deep blindness as an ideal is not universally shared by those who become blind later in life: Other journeys into

blindness have featured the heightening of visual memory and visual consciousness.[1] This divergence illustrates the point in question: It is not the loss of sight per se that transforms these people, but the way in which they react to and grow into their new, sightless, life. Becoming blind can, at least for some people, be transformative activity.[2]

It may be that the right way to understand becoming a parent or becoming blind is as involving both transformative revelations and transformative activities; or it may be that these changes are more like revelations for some and more like activities for others. But the question is an intelligible one in any case: It makes sense to ask, is such-and-such a transformation a revelation or an activity? And the answer makes a big difference: Coming to taste durian fruit and coming to taste the note of cherry in wine are very different sorts of transformations. This distinction seems to mark a real difference. But what exactly is it that differentiates a transformative revelation from a transformative activity?

2.1. Temporal Profiles

The most obvious difference is temporal. Hull's journey into deep blindness extends over years. If one must learn and grow into parenthood, that will take years or perhaps decades. Learning to taste the notes of cherry in wine may take weeks or months. Transformative revelations are comparatively speedy: being bitten by a vampire, feeling one's unborn baby kick, tasting durian fruit, seeing color for the first time—all of these events happen in less than a minute.

I do not think we can ground the distinction solely on speed. Granted, it is hard to picture a human being completing Hull's transformation in seconds. But that seems to be a feature of human psychology and physiology, not a feature of the structure of the activity. These activities essentially take time, but they do not essentially take years as opposed to minutes. Nor is it true that transformative revelations are, of necessity, sudden. Imagine a slow-acting vampire poison: the vampire bites you, and one month later you (slowly) develop an aversion to sunlight, the next month a taste for blood, the next month your senses sharpen, etc. You become a vampire over the course of a year. Even though you come to know what it is like to be a vampire slowly, over the course of a year, the transformation remains a revelation.

Why? We might suppose that the explanation is as follows: Over the course of that year, you are not working to learn to become a vampire. Becoming a vampire is something that happens to you, rather than something that you do. This suggests that perhaps the difference between the two kinds of transformation corresponds to that between agency and patiency.

[1] See Sacks (2003).

[2] Likewise becoming sighted, the problematic nature of which is the subject of a large literature.

2.2. Active vs. Passive

The proposal is as follows: In transformative revelations, you are transformed by what happens to you, whereas in transformative activities, you yourself do the transforming.

Ruth Chang observes that in the cases described by Paul and Ullmann-Margalit, the fact that the person consented to the transformation is incidental to the transformation that takes place. Consider the vampire example. What does the transformative work is the event of being bitten; the bite transforms a person in just the same way regardless of whether she has consented to it. She calls these "*event-based transformative choices*." Let us explore whether Chang's contrast class—what she calls *choice-based transformative choices*—corresponds to transformative activities.

In *choice-based transformative choices*, "the making of the choice itself transforms you." Chang holds that by performing the very act of choosing to have a child (as opposed to remaining childless), or choosing to vacation on the beach (as opposed to the mountains), you change yourself into a person who has reasons she didn't have before:

> ... suppose you commit to some feature of the beach vacation, thereby making it true that you have most reason to go on the beach vacation. You change "who you are" in this small way by creating for yourself a new reason you didn't have before. You are now, to that small extent, a beach person rather than a mountain person. (2015: 280)

It may be that sometimes the very act of choosing generates reasons to follow through—this certainly seems true of cases in which the choice coordinates the activity of multiple people. This does not, however, seem to be what is happening in a transformative activity. The transformation into someone who grasps a reason to e.g. choose this wine on this occasion is not made simply by choosing to become a wine expert.

Likewise, Hull did not arrive at deep blindness by decision, but rather over the years of reflecting, conversing, and writing about what he was experiencing. Did he preface this reflecting, conversing, and writing by deciding to become deeply blind? No: because his grasp of deep blindness was itself a product of reflecting, conversing and writing on his condition. But even if he had, that choice couldn't have transformed him in the relevant respect. Suppose a newly blind person, influenced by Hull's book, decides to become deeply blind. That decision does not, by itself, transform her into someone who is deeply blind. It may, if Chang is right, give her reasons to become

---

Ackroyd etal. (1974) describe a patient who chooses to return to living as a blind person after having her sight restored: "'Seeing,' far from being a rewarding activity, had become a tiresome duty for her, and left to herself she soon lost interest in it" (p.116). Thanks to John Schwenkler for directing me to this material.

deeply blind, but the relevant change is yet unaccomplished at the time of decision. She doesn't become deeply blind just by deciding to.

Let us, then, set aside choice-based transformative choices. Could there be another way to use the distinction between agency and patiency to distinguish transformative revelations from transformative activities? Consider the following proposal: In the case of transformative revelations, agency is restricted to an initial moment of decision, whereas in the case of transformative activities, the subject is at work throughout her transformation. After presenting her neck to be bitten, or placing the durian fruit in her mouth, or having unprotected sex, the agent has only to sit back and wait for the transformation to take place. Her becoming a vampire, or acquainted with the taste of durian fruit, or pregnant, is not, after this point, her own doing. It is something that happens to her.

In a transformative activity, the agent is at the helm of her transformation for its duration. If, at any point during her transformation, she stops doing what she is doing, her transformation likewise comes to a stop. Moreover, she controls the form it takes: The way in which she is transformed is guided by and responsive to her activity. Consider someone learning to become a music lover. If the agent drops out of her music appreciation classes, stops listening on the way to work, and cancels her concert subscription, her musical development will come to a stop as well. She will thereby put a stop to the process in which she develops new musical tastes, comes to appreciate new features of music she has already heard, and forms new music-related desires. Becoming musically transformed is not something that happens to you—it is something that you do. If you stop doing it, it stops happening.

In a transformative activity, the transformation is conditional on activity throughout, rather than being independent of their activity after some initial stage. Call this the "activity-dependence" criterion. Can we use this criterion to distinguish transformative revelations from transformative activities? I don't think so, for we can construct a transformative revelation that is activity-dependent. Let me, once again, ring some changes on the vampire story. Imagine that, in addition to being slow- acting, the vampire toxin is also one that can be counteracted by what the bitten person does. So, for instance: If she wears sunscreen she will never develop sunsensitivity; if she becomes a vegetarian, she never develops a taste for blood; if she makes affectionate physical contact with a non-vampire she never attains immortality. In order to become a vampire, such a person would need not only to wait for the toxin to take full effect, but also to cooperate: eating meat, avoiding sunscreen and hugs. So long as each of the relevant changes has a counterpart activity, the transformation will be thoroughly activity-dependent. The agent's input will not be restricted to an initial moment of consent, but extend throughout the transformation.

And yet, I maintain, becoming a vampire continues to be a transformative revelation rather than a transformative activity. Even though the revelation has been spread out over time, and regulated by the person's consent throughout, it is not active in the relevant sense. But what exactly is that sense? In a transformative activity,

what the agent does to transform herself is to reach out and grasp after the value or experience that is the target of the activity. Hull's agency—his reflection, and talking, and writing—is directed at learning what blindness is like, and what is good about it. Likewise, the person who is learning to appreciate music or wine is acting so as to understand the experience or value at the heart of the relevant activity. This might lead us to place learning at the heart of transformative activity. Can we differentiate transformative revelation from transformative activity by using learning as a criterion?

2.3. A Learning Activity

Transformative activities are learning processes: Learning how to be blind, or what the virtues of wine consist in, or why one should listen to classical music, or what it is to be a parent. But a transformative revelation is also a learning process. The eater of durian fruit, or newly transformed vampire, comes to know something she did not know before—what it is like to be a vampire or taste durian fruit. And the reason she initiates this process (and, if necessary, sustains it) is precisely that she wishes to acquire the relevant kind of knowledge.[3] So both kinds of transformation are learning processes.

All transformative experiences involve entering into a condition in which one knows what it is like to e.g. care about one's children, taste a new food, enjoy music, be immortal. At the end of every transformative experience, the agent has learned something. But consider two different forms that learning can take. In a transformative revelation, what the agent does to facilitate the learning is not itself a case of learning. A transformative activity, by contrast, takes the form of inquiry. Every part of it is learning. It is the difference between learning by being informed and learning by figuring out.

We are now in a position to draw the relevant distinction between transformative activity and transformative revelation. Both kinds of transformative experience satisfy these two criteria:

> (1) **The Agency Criterion**: the person chooses or enacts or sustains or engages in the transformation. She transforms herself.[4]
>
> (2) **The Learning Criterion**: the person acquires knowledge of what something is like. She comes to have a new experience.

In a transformative revelation these criteria correspond to separate elements of the story. The agent's doings and her comings-to-know are not identical. So, for instance,

[3] Or so Paul maintains: "[In transformative choices] we choose between the alternatives of discovering what it is like to have the new preferences and experiences involved, or keeping the status quo" (2014:122).

[4] Because both Paul and Ullmann-Margalit ask the question of what it would be to rationally

what she does is submit her neck to be bitten, or avoid hugs. Those actions are not themselves identical to her (growing) awareness of what it is like to be a vampire. Her disposal of sunscreen and vegetables, or retreat from an oncoming hug, are not themselves learning experiences. They only facilitate the eventual, or perhaps concomitant, grasp of vampiricity that the toxin wreaks upon her. What she did and what she learned were different. She acted in order to learn. Her acting was not itself learning.

In a transformative activity, the doing and the learning are not distinct from one another. Whatever the agent was doing in order to bring it about that she become X-ish was identical to her learning to be X-ish. This is why the transformation—her coming to learn what music, or deep blindess is like—must stop if she stops working. Her doing is her learning.

Paul and Ullmann-Margalit have explored the question why should one perform the action component of a transformative revelation: why should one have unprotected sex, or present one's neck for biting, or eat a durian fruit. Notice that, in these cases, there is no attendant question about how one does these things. One can successfully put a durian fruit into one's mouth, get bitten by a vampire, or have unprotected sex without knowing what one is getting out of those actions. The separation between the action component and the learning component ensures that performing the relevant action doesn't itself call for the knowledge the person is trying to acquire.

In a transformative activity, the "how" question looms just as large as the "why." When one is learning by doing, what one is learning is also what one is doing. This means that transformative activities involve doing what one does not yet know how to do, for reasons one does not yet grasp.

Consider: How does one become a wine connoisseur, or deeply blind, or a classical music aficionado? In each case, one does so by performing the relevant activity: tasting wine, not seeing (at all!), listening to music. The problem is that these activities represent both means and ends. How am I to come to listen to music—to really listen—if that involves (really) listening to music? It would be different if I were in some kind of vampire scenario, and someone said to me, "Push this button and you will be injected with a serum that turns you into a music lover." I might, in the manner of Paul and Ullmann-Margalit's agents, agonize over whether to push the button. But I at least know how to push a button: You just move your finger onto the button and press. What I do not know how to do is (really) listen to music. I can sit there while the music is playing, but that's not the same thing as really listening for the right things, in the right way. Likewise with wine-tasting. I learn to taste by tasting.

Transformative revelations require me to do something for the sake of an end I do not yet grasp; transformative activities require me, in addition, to do so by way of an activity I do not know how to do. A number of questions arise about the intelligibility of acting without knowing what I'm doing nor why I'm doing it. First, about the ratio-

choose to become transformed, they set aside those cases in which a person has no say in becoming transformed.

nality of such activity. If one's agency extends all the way to the new condition, one must have some kind of reason for pursuing that condition. What kind is it? Second, there are questions about the moral psychology of transformative activities. The agent of such an activity finds herself conflicted between engaging in an activity whose point she doesn't understand, and falling back on the many goal-oriented activities that do make sense to her. How does she motivate herself to continue? Finally, transformative activities raise metaphysical questions as to how a person can make herself into something she is not: How does the theorist of transformative activity avoid the paradoxes associated with the concept of self-creation?

I discuss these questions in Callard (2018); here I raise them only as indications of the sui generis character of this kind of transformative experience. In what follows, I will not attempt to defend but rather to illustrate the concept of transformative activity.

3. Elena and Lila

3.1. Another Kind of Example

The examples of transformative activity given above have all been: Examples of (a) adults transforming themselves into people who differ (b) in relatively superficial ways from the people they were before while located (c) in environments hospitable to the transformation. If a musical appreciation or viniculture class is on the table as an option for you, the world is going out of its way to pave your transformative path. I want now to consider a case of transformative activity in which (a) children and teenagers are (b) radically remaking their identities (preferences, values, experiences) in an environment that (c) actively resists those efforts. In this second kind of case, the person is likely to have such an attenuated grasp of where she is going that we have trouble seeing her as headed there at all.

The simpler examples of transformative activity described in the first, taxonomical part of this chapter are helpful in clarifying the concept in question; but the messier case to which I am about to turn is better able to showcase the power and promise of this bit of conceptual machinery. Nascent rationality is often misclassified either as irrationality or as non-transformative rationality. We tend to feel that we must choose between understanding someone as passively transformed by what happens to her or as antecedently committed to the (putatively newly chosen) values and experiences. Either she is changed or she doesn't (really) change. The concept of transformative activity gives us a way to read the messy and misdirected moral psychology of the young, confused learner as a window into the phenomenon of radical self-transformation.

3.2. Competitive Friendship

I return to Ferrante's *My Brilliant Friend.* At the age of 5 or 6, Elena and Lila often play with their prized possessions—two dolls—side by side, each girl pretending to ignore the other. Their friendship begins not at the moment they decide to share dolls, but right afterwards: as soon as Lila gets her hands on Elena's doll, she throws it down into a dark cellar. Elena doesn't express surprise, or cry, or protest. Instead, she does the same to Lila's doll. "What you do I do," says Elena. This exchange sets the tone for a lifelong rivalry.

The competition between Elena and Lila structures their relationship from start to finish, serving as the narrative frame of the novel: Elena, in her 60s, discovers that Lila has vanished and eliminated every trace of herself in her home, to the point of cutting herself out of photos. "She wanted not only to disappear herself, now, at the age of sixty-six, but also to eliminate the entire life that she had left behind. I was really angry. We'll see who wins this time, I said to myself. I turned on the computer and began to write—all the details of our story, everything that still remained in my memory" (Ferrante 2012: 23).

Their competitiveness often turns so bitter and hostile as to verge on sadism. At one point, Lila, whose parents forbid her from having a higher education, tries to sabotage Elena's chance at one. At another point, Elena considers sabotaging Lila's marriage. They steal one another's lovers and belittle one another's achievements. Elena describes an emotional scene in which their elementary schoolteacher expresses regret to Elena over not having pushed harder to secure Lila an education, "as if the teacher were realizing that something of Lila had been ruined because she, as a teacher, hadn't nurtured it well. I felt that I was her most successful student and went away relieved" (Ferrante 2012: 277). And yet, although each of them goes on to marry and have children, the relationship between the two women remains the central fact of each of their lives. And this not in spite of but because of the element of antagonism between them, which turns out to be the fundamental source of creativity and vitality for each: "I soon had to admit that what I did by myself couldn't excite me, only what Lila touched became important. If she withdrew, the things got dirty, dusty" (Ferrante 2012: 100).

Intense competition—especially between women—characteristically elicits both salacious interest and facile opprobrium. It functions as foil to the supportive, mutually affirming relationship that standardly serves as a model for female friendship. Elena and Lila's passionate, lifelong antagonism may even strike the reader as a pathological or a diseased form of human relationship. I want to try to show you that we can see what is going on between them differently if we look at it through the lens of transformative activity.

And we should want a different way of looking at them. As I mentioned at the opening and will go on to discuss in more detail, Elena and Lila's competition manages to bear extraordinary fruit in an otherwise barren landscape: It fuels their creative, polit-

ical and intellectual endeavors in a community that pushes them to be anything but creative, political, and intellectual. If pathology is defined in terms of the impediment to human functioning, it has to be admitted that, on balance, their relationship serves the development and actualization of their capacities more than it hinders them. So it would be good to have a theory of how and why this is so.

I propose that their competition provides both the motivation and means for pursuing goals that would otherwise be outside the girls' field of view. Competition answers the "how" question: Do what Lila is doing, only better. And it answers the "why" question: Do it because Lila is doing it. Competing with someone is not the only way of doing what you do not know how to do,[5] but the claim I will explore in the rest of this chapter is that it is a way of doing so. Each girl is, for the other, the answer to the question: How does one transform oneself when one lacks the resources to do so? How does one do what one does not yet see as good, for a reason one cannot yet appreciate?

3. Aspirational Competition

When Lila loses interest in competing against Elena academically, the nature of Elena's academic interest changes as well: "since Lila had stop pushing me, anticipating me in my studies and my reading, school ... had stopped being a kind of adventure and had become only a thing that I knew how to do well and was much praised for" (Ferrante 2012: 187).

It is important to distinguish what I will call the "aspirational competition" that characterizes Lila and Elena's relationship from two other forms competition can take. When Lila loses interest, Elena does not stop competing academically. Nonetheless, that competition devolves into something more instrumental. She competes in order to secure a good—praise—to which competition happens to be the means. She competes not in order to learn how to do something new ("a kind of adventure"), but because there is something she already "knew how to do well." We instrumentalize competition when there is some good that we want, and, as it happens, we must compete to secure it. If, for instance, you and I both want the same job, we may have to compete for it. It is a mark of the non-aspirational character of this competition if I respond in a purely positive way to the discovery that you, having been awarded an even better job, have withdrawn from the contest.[6]

Aspirational competition is also distinct from the competition that is part of intrinsically adversarial undertakings, such as sports (teams or individuals compete against one another), elections (candidates competing against one another), and courtroom trials (prosecution competing against defense). It is an intrinsic feature of those activities that success entails surpassing another. That feature is strikingly absent from e.g.

[5] In pp. 135-42 of Callard (2016), I discuss six other ways.

[6] For a discussion of the broader category of non-aspirational phenomena into which instrumentalized competition fits—I call it "self-cultivation"—see Callard (2018: ch. 1, pt 3).

the activity of learning ancient Greek. It is not competitive in the way that running a race is competitive.

If we think back to the most competitive periods of our lives, our minds are likely to settle on some part of our schooling. Why is school such a competitive place? The goal of becoming educated is, on the face of it, totally unrelated to the goal of surpassing one's classmates. And yet many of us find it natural to approach any group learning environment—be it a math class or a yoga studio—in a competitive spirit. In the moment, it feels to the learner as though what really matters is getting the highest grade on the math test, or lifting one's leg a fraction higher than one's neighbor. But it is not satisfying to get the highest grade by some oversight on the teacher's part, or if the height of your leg is due to the fact that you are standing in an elevated spot. We want to *really* beat our competitors—which is to say, to beat them in such a way that our victory is a sign of our excellence at math or yoga. But why, in that case, don't we go straight for the brass ring, and aim at excellence in math or yoga irrespective of whom we "beat" in the process? The answer must be that, as newcomers to the appreciation of math and yoga, we have not yet transformed ourselves into the people for whom that brass ring is squarely in view. In those cases, it is by way of our competitors that the relevant form of excellence shows up as an object of pursuit.

The choice of competitor reflects its aspirational character: In a math class, you might set yourself against the best student, so as to become as good as possible. But if she happens, unlike you, to be a mathematical genius, you might find yourself ignoring her scores and competing against someone in your league. One also chooses competitors on the basis of one's own interests and passions—if math is not my subject, I might be relatively uncompetitive in that class, but fiercely so in music or science. You are competing against whoever will help you become better, in whatever you (are coming to) want to become better in. Insofar as you engage in aspirational competition, you see your competitor as a rope by which you pull yourself up by your bootstraps. She is the door in the valuational room that would otherwise trap you into being stuck as you are.

4. The Ogre of The Neighborhood

Consider the event that cements Lila and Elena's friendship. Early in the novel, the two girls make the unthinkable decision to confront the "ogre" of the neighborhood, the terrifying Don Achille. They have been forbidden by their parents from speaking to or even looking at him: "regarding him there was, in my house but not only mine, a fear and a hatred whose origin I didn't know." Convinced that Don Achille has taken their dolls from the basement into which each threw the other's, Lila and Elena ascend to his apartment to demand their return. Elena reports: "I had forgotten every good reason, and certainly was there only because she was. We climbed slowly towards the

greatest of our terrors of that time, we went to expose ourselves to fear and interrogate it" (Ferrante 2012: 29).

When Elena speaks of herself in the singular ("I had forgotten ... "), she sees no reason for proceeding. Her point of view makes confronting Don Achille an impossibility: "I thought that if I merely saw him from a distance he would drive something sharp and burning into my eyes. So if I was mad enough to approach the door of his house, he would kill me." When she thinks for herself alone, approaching Don Achille is not an option on her deliberative horizon. But when she adopts the "we" together with Lila ("we climbed," "we went to expose"), she gains the ability to move towards what she sees as certain death: "At Don Achille's door my heart was pounding, I could hear it in my ears, but I was consoled into thinking that it was also the sound of Lila's heart" (Ferrante 2012: 65).

Together, they summon the courage to violate the law of the neighborhood, and overcome the fear that is their cultural inheritance: "We were forbidden to go to Don Achille's, but she decided to go anyway and I followed. In fact, that was when I became convinced that nothing could stop her, and that every disobedient act contained breathtaking opportunities." Lila leads the way, and the need to keep up with her makes it possible for Elena to decouple her own fear from the attendant impulse to flee. The girls become able to feel fear without being driven by it, and this puts them in a position to look at the fear, to "interrogate it."

How do the girls secure this distance from their fear? One might suggest that Lila constitutes, for Elena, a symbol of the possibility of freedom from the neighborhood. But I think this way of putting their relation is too abstract and etiolated. In order for Lila to symbolize freedom, Elena would have to have an independent grip on freedom, and then see Lila as that. Lila is closer to a conduit than a symbol. She doesn't merely represent "breathtaking opportunities"; rather, engaging with her constitutes the opportunity in question.

In order to be working towards her future self, Elena must, in some sense, be in conversation with the person she will become. In the next section, I will explain why her relationship with Lila is the form that that conversation takes.

5. Competition as Escape

To be such as to care passionately about being the best at math or yoga is to already be attuned to potential transformations in those areas. It is a mark of privilege to have those possibilities present themselves—even if they do so only opaquely, in the form of your inclination to compete against a fellow student. A person who is inclined to aspirationally compete over math or yoga has already been, to some degree, set up for transformative activity. She has been raised to see that there are avenues of value available for her to explore. Elena and Lila live in an impoverished neighbor-

hood of postwar Naples: their childhoods are structured by violence, sexism, and the expectation that they will walk the paths trodden by their own mothers.

When, against tremendous familial and cultural pressure, Elena manages to complete middle school, she assumes she will begin working. Her teacher insists, "you have to go on studying." Elena is genuinely perplexed:

> I looked at her in surprise. What was there left to study? I didn't know anything about the order of schools, I didn't have a clear idea what there was after the middle school diploma. Words like high school, university were for me without substance, like many of the words I came across in novels. (Ferrante 2012: 123)

Elena doesn't, at this stage, have a way of grasping the value of an education—she doesn't grasp that it is her way out of the neighborhood. But she *does* see Lila that way: "As a child I had looked to [Lila], to her progress, to learn how to escape my mother" (p. 322). Elena's competition with Lila is about escaping into something she can only get the faintest glimmer of, namely the possibility of living a life unlike the only one she has been presented with:

> something convinced me, then, that if I kept up with [Lila], at her pace, my mother's limp, which had entered into my brain and wouldn't come out, would stop threatening me. I decided that I had to model myself on that girl, never let her out of my sight, even if she got annoyed and chased me away. (Ferrante 2012: 46)

Lila, like Elena, "was struggling to find, from inside the cage in which she was enclosed, a way of being, all her own, that was still obscure to her" (p. 295). Each girl yearns to escape the strictures of the world into which she has been born. Lila, whose parents forbid her from post-elementary schooling, aims to do so by remaking her environment: "Did she (Lila) want to leave the neighborhood by staying in the neighborhood? Did she want to drag us out of ourselves, tear off the old skin and put on a new one, suitable for what she was inventing?" (p. 46). Elena, by contrast, leaves the neighborhood by leaving the neighborhood, and higher education—at the end of My Brilliant Friend, she is headed to university—is her ticket out.

The transformative bond between Lila and Elena is the object of Ferrante's attention throughout the novel:

> There was something unbearable in the things, in the people, in the buildings, in the streets that, only if you reinvented it all, as in a game, became acceptable. The essential, however, was to know how to play, and she and I, only she and I, knew how to do it. (Ferrante 2012: 107)

Together, they begin to see the alternative possibilities—alternative values, activities, ways of being—that constitute, for each girl, a way of life. Her relationship with Lila strikes Elena as a game which allows them to escape the strictures of their environment. The idea that they are playing a secret game is a way of inchoately grasping the thought that there could be other rules to life than those they have been taught. The game is a world inside the world; it is a proxy for the world outside the world—the world outside the neighborhood—for which Elena is only beginning to learn how to yearn. The goal of being and beating Lila is the closest Elena can get to the idea of the value of escape, just as getting a higher grade on a test can be the closest that you get to the idea of mathematical excellence.

If we ask, "Why couldn't Elena straightforwardly work to become different from her mother?", we are failing to appreciate the difficulty of seeing one's way out of the only value-system to which one has been exposed. The strictures of the world into which one has been born are, at the same time, the strictures of one's own mind. Seeing one's way out of that box is a feat that Elena accomplishes with Lila's help; it is no wonder that Lila strikes her as magical:

> ... she took the facts and in a natural way charged them with tension; she intensified reality as she reduced it to words; she injected it with energy ... as soon as she began to do this, I felt able to do the same, and I tried and it came easily. (Ferrante 2012: 130)

Consider the words Elena associates with Lila in the passages I have quoted here: "power," "energy," "tension," "opportunity," "intensified," "progress," "adventure," "important." Each of those words is a promissory note. Engaging with Lila allows Elena access to progress without (fully) knowing in what dimension she is progressing; it allows her to feel a sense of adventure without knowing where the adventure is headed. Her relationship with Lila is a partial glimpse into the future—one that makes it possible for the cosmopolitan, feminist novelist that Elena becomes to be the work of the clueless, impoverished child who becomes her.

Aspirational competition of this kind constitutes a kind of transformative activity. It is a way of learning how to be different by acting differently. Such a competitor isn't working to beat her antagonist for its own sake, as in intrinsically adversarial activity. Nor is competition for the sake of something she independently wants, as in the case of instrumentalized competition. The aspirational competitive impulse is directed at a target one doesn't yet know how to want. In this kind of competition, beating someone is becoming someone:

> I, I and Lila, we two with that capacity that together—only together—we had to seize the mass of colors, sounds, things and people, and express it and give it power. (Ferrante 2012: 138)

References

Ackroyd, C., N. K. Humphrey, and E. Warrington. 1974. "Lasting Effects of Early Blindness: A Case Study." *Quarterly Journal of Experimental Psychology* 26(1): 114-24.

Callard, A. 2016. "Proleptic Reasons." *Oxford Studies in Meta-Ethics* 11: 129-54.

Callard, A. 2018. *Aspiration.* Oxford: Oxford University Press.

Chang, R. 2015. "Transformative Choices." *Res Philosophica* 92(2): 237-82.

Ferrante, E. 2012. *My Brilliant Friend,* trans. Ann Goldstein. New York: Europa Editions.

Hull, J. 1990. *Touching the Rock: An Experience of Blindness.* New York: Penguin Random House.

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

Paul, L. A. 2015. "What You Can't Expect When You're Expecting." *Res Philosophica* 92(2): 1-23.

Sacks, O. 2003. "The Mind's Eye: What the Blind See." *New Yorker* (July 28): 48-59.

Ullmann-Margalit, E. 2006. "Big Decisions: Opting, Converting, Drifting." In A. O'Hear (ed.), *Political Philosophy,* 157-72. Cambridge: Cambridge University Press.

9. Transformative Expression[(12)]

Nick Riggle

1. Introduction

It's common sense that art can change our lives and selves: novels, poems, films, plays, music, operas—they affect us in profound and transformative ways. Artists needn't intend to have such an effect on their audience; they aim to produce something of value, something worthy. But it turns out that aesthetic value is like that; it has that power. It can change us.[1]

Art can change our lives, but can it be designed to do so directly?[2] Must we visit a museum to look at grand and expensive paintings, or can art confront and change us in the course of everyday life? If so, then how far can we go? Can we change not only individual lives, but communities or even entire societies? Many twentieth-century artists answered these questions in the affirmative, and their affirmation found expression in a vast range of ingenious and ambitious postwar, postmodern, and contemporary works: "The most radical theses of the European avant-garde during the revolutionary upheavals of 1910-1925: that art must cease to be a specialized and imaginary transformation of the world and become the real transformation of lived experience itself" (Clark et al. n.d.). Over the following century these "radical theses" informed a range of inventive and ambitious works that aim to directly engage people in the course of their lives and move them to express themselves in transformative ways.

[1] I discuss the transformative power of beauty in Riggle (2016). For a discussion of morally transformative art, see Walden (2015).

[2] My focus here is on a specific tradition in European and Anglo-American art (and various movements around the globe influenced by or responding to these traditions), but it is important to note that in some aesthetic traditions the answer is an obvious "yes." Consider this passage from Robert Carter: "The practice of a Japanese art is in all respects transformative. Each art is designed to make one a different person, a better person" (2008: 4). I'll be interested in what follows in how exactly transformative art is supposed to make us "better." Thanks to Julianne Chung for calling my attention to this passage. For more on the topic, see Chung (2018).

(12) Nick Riggle, Transformative Expression In: Becoming Someone New: Essays on Transformative Experience, Choice, and Change. Edited by: Enoch Lambert and John Schwenkler, Oxford University Press (2020). © Nick Riggle.
DOI: 10.1093/oso/9780198823735.003.00010

But there is a whiff of paradox about this aim. How can we express ourselves in a way that transforms the self we express? We tend to think that a person engages in authentic, or exemplary self-expressive, action only when the action issues from, speaks to, or otherwise embodies their core commitments. A necessary condition of such expressive action, the thought goes, is that it must be expressive of a self—one's true, authoritative, or authentic self—embodied in the preferences, rules, beliefs, desires, or values that the person "owns" and that issue, motivate, or endorse the action. And the action is more authentic the more central to the self those values are.[3] Such a view is suggested by Sartre's example of the waiter whose actions issue from his sense of what a "waiter" should do but without any substantial commitment to being a waiter. Sartre claims that such a person is acting in "bad faith" precisely because he is not acting from an accurate, or even any, sense of who or what he is. Likewise, when we are forced to act a certain way by social or professional norms, bad jobs, peer pressure, tradition, and so on, we often think we aren't acting well, precisely because our actions fail to express our plans and values, what we really care about, or who we really are. Some philosophers extend such a necessary condition beyond authenticity to autonomy. Your actions are not self-governed, according to such views, unless they embody your plans, the rules or values you endorse, your commitments and concerns, and so on.[4]

More generally, philosophers select some privileged set of dispositions—acting on one's cherished values, higher-order desires, plans, moral convictions, weightiest reasons, etc.—and say that one's actions are authentic or autonomous only if they express or issue from these dispositions. Call this family of theories dispositioncentered theories of self-expression:

> If A's ^-ing is authentic/autonomous, then A's ^-ing expresses A's selfconstituting dispositions.

But if self-expression is an achievement that is the expression of self-constituting dispositions, then how could expressing oneself change those dispositions? The question is especially pressing when the relevant action is one the transformed agent would have rejected, would not have endorsed, desired, sought out, or performed, prior to the effect it has on that very agent. Yet that is precisely what is envisioned by so many avant-garde artists who sought to dislodge or replace people's core commitments in a way that they would embrace. The participants in these artworks would seem to be expressing a self they neither have nor want by way of adopting a new one. How could an artwork do that, even in theory?

In what follows I develop the concept of "transformative expression" and argue that: (1) transformatively expressive acts feature in a range of avant-garde artworks that

[3] A classic statement of such a view is Frankfurt (1971).

[4] For discussion of the relation between authenticity and autonomy, see Oshana (2007).

confound standard ways of thinking about aesthetic value, and (2) they are counterexamples to disposition-centered theories. These two points turn out to be related: expanding our understanding of aesthetic value to include a range of "playful" actions shows how we can express ourselves even when we are not expressing our commitments. If this is right, then standard theories of aesthetic value and common ways of thinking about authenticity and autonomy should be rejected for the same reason. I sketch better ways of thinking about both.

2. Defining Transformative Expression

We can begin to see a problem with the standard way of thinking about selfexpression by considering ways of coping when, for one reason or another, we cannot act on our plans, values, or commitments. There was a solid several months in my recent past when I wasn't myself. I hadn't stopped existing, entirely, but I was on the academic job market. I was teaching too much at a small liberal arts college in a cold place with few friends around, uncertain prospects, and no discernible desire to continue. I found myself in a kitchen cooking. A lot. Elaborate fish dinners for midweek lunch, various things pickled or preserved, and so many kinds of squash. Did I really want to do this? I think I would have preferred to be wholeheartedly engaged in philosophical activity—the activity around which many of my core values and desires revolved. I wondered, Who does this? Who cooks meals like this for lunch on a Wednesday? Bored chefs maybe. Not me. But it seemed like the thing to do, or at least a thing to do, when the thing to do could not really be done.

Sometimes we do things without being able to relate what we are doing to our normal sense of self—because we are lost, depressed, exploring, ambivalent, curious, desperate, or just compelled to do *something*—and sometimes expressing some self or other in this way is enough to transform us into some self or other. I now consider myself a pretty good cook.

The exploration of cooking was suggested to me, in part, by the absurd number of shows I was watching on the Food Network and the Cooking Channel. *Restaurant Impossible*, hosted by Robert Irvine, is about redesigning restaurant interiors, often on a strict budget. The *Restaurant Impossible* design team puts together some impressive designs, but they usually aren't very original—many of the motifs, themes, and much of the furniture is seen all over the United States, very likely in the same towns and cities as the featured restaurant. No doubt the restaurant owners have seen and ignored, or maybe even dismissed, such designs. Yet almost without fail (it's a TV show, after all), when the owners see their redesigned space full of these mostly familiar design elements, they break down in tears of amazement, joy, and deep gratitude. They see themselves anew in these designs and their lives as restaurant owners are changed.

People on "makeover" shows like TLC's *What Not to Wear* or Netflix's *Queer Eye* have a similar reaction when they get a new haircut and a few new outfits—hairstyles

and outfits that they no doubt have seen multiple times, maybe even worn once or twice, and likely rejected as something they would never wear. But when they turn around and look in the mirror and at their friends, they seem to be transformed.[5]

Let's call these events "transformative expressions" because they are expressive acts that transform the individual. Here's a way to make the idea I have in mind more explicit. For a person N, an action-type A, and a time interval *t,* N's doing A during *t* is transformatively expressive if

1. Prior to t: ^-type action is not endorsed by N's core commitments.

2. After t, ^-ing expresses N's core commitments.

3. (2) is a direct result of N's ^-ing during t.

Apply this to a makeover case. At and prior to getting the makeover, Dagny does not want to change her personal style—she's committed to it, decidedly happy with her '80s haircut and shoulder-padded blazers. After the makeover, Dagny very much wants to style herself that way—so much so that it seems to her to be one of the more meaningful things she can regularly do for herself. And this is a direct result of acting in a way that is contrary to her sense of self and style. She wouldn't feel this way unless she took the plunge and tried a new look.

The interval t might be short and involve one performance of φ-type action, but it might be extended across repeated performances of φ-type action. Someone who dislikes church might think it's their only option (perhaps they're persuaded by Pascal's Wager). They go through the motions of attending church every Sunday— singing the songs, participating in the prayers, listening attentively to the sermons— and after an interval of attendance, going to church is expressive of their core commitments. (The example also shows that φ-type action, e.g. going to church, can consist in multiple action-types.)

We can imagine a case where (1) and (2) are true but not (3). Imagine that Dagny's new style is not a result of her adopting a new look but rather a result of a new dress code at work that requires proper business attire. She adopts the way of dressing in order to conform to work standards and not be fired; over time, and for reasons other than her conforming to the dress code, she personally changes—she becomes more of a team player, more confident, more organized. Now her core commitments are compatible with, and even expressed through, dressing in proper business attire, but not as a direct result of her (even repeatedly) adopting an unfamiliar look. So dressing that way is not a transformative expression.

[5] There are other intuitive examples: A person in therapy might express her feelings or voice judgments, and in doing so come to own those feelings or judgments in a transformative way. The emerging literature on games and practical reason emphasizes how games facilitate our trying on different deliberative perspectives and elicit experiences of freedom in potentially transformative ways. See Nguyen (forthcoming); see also Gingerich (2018).

I don't mean anything too fancy by the vague notion of a core commitment. A "commitment" here is a stand-in for a range of attitudes. I might be committed to a person because I deeply admire him; I might be committed to a certain artwork because I have a meaningful attachment to it; I might be committed against cucumbers because I really dislike them. A core commitment is one that has a certain importance for the person and is a matter of degree. What makes a commitment core is that it partly constitutes one's individuality—our core commitments are a large part of what makes us what we are like, guiding what we do with our lives and how we do it: what we value, appreciate, avoid, or seek out; the social roles we identify with; the kinds of people we want to be around. A core commitment is such that losing it would require us to redraw how we think about who we are and what we care about.[6] And we might do that a fair amount over time—I don't assume that our core commitments answer to or pick out a "true" self or a fixed self. I have various "commitments" that are not core: I avoid cucumbers, prefer not to bowl (because no matter how hard I try I'm really bad at it), and am reliably down to go on a nice long walk. Changes in these preferences would not change my individuality or threaten my sense of self. Sometimes we do things that are incompatible with our peripheral commitments. I'll throw some cucumbers on a sandwich even though I don't really like them. Maybe one day I'll think, "You know, cucumbers aren't so bad after all." This isn't transformative because the commitment isn't core. I've simply changed some of my peripheral but reliable preferences.

On this way of construing things, the action mustn't be endorsed by N's core commitments. The comedic Parks and Recreation character Ron Swanson—a passionately pro-America, anti-government libertarian—regards going to Europe as incompatible with his core commitments. When he finally goes to London, he's basically vindicated. But then his awesome friend Leslie Knope arranges for him to visit the Lagavulin distillery—his favorite whisky—and he realizes that places in Europe really do satisfy his core commitments. This is not a transformative expression so much as a kind of awakening. Ron's commitments don't change, but he's awakened to a wider world in which they can be exercised. (Note that it's not quite right to say that he has a new core preference for visiting Europe. Perhaps he does in a sense, but only insofar as it speaks to his unchanged commitment to Lagavulin whisky.) Of course, transformative expressions do not always transform you into a better person.

So what is the relation between transformative expression and transformative *experience?* L. A. Paul (2014) focuses on transformative experiences that are both "epistemically" and "personally" transformative (p. 17). An epistemically transformative experience introduces you to a new phenomenal content—the taste of amaranth, the

[6] My notion of "core commitment" is the basically the same as Chesire Calhoun's (2009) notion of a "substantive commitment" except "core" is alliterative and only has one syllable. Calhoun writes, "The commitments I have in mind are ones whose objects are candidates for inclusion in a life plan, or that give shape to a life, or define an identity, or answer the question of what one's life is about. Intuitively, sexual, ethnic, and religious identities, place of geographic residence, avocations and careers, and friendships and intimate relationships would count as such candidates" (2009, p. 614).

sound of a steel tongue drum. A personally transformative experience substantially changes your point of view, by changing either your core preferences (p. 16) or how you experience being who you are (p. 16). But paradigmatic transformative experiences in this sense are not transformative *expressions* in the sense defined here. Consider becoming a parent. For many prospective parents, being a parent is endorsed by their core commitments, so condition (1) of the definition of transformative expression is not met. For many prospective parents, becoming a parent is something they really want, something they hope and prepare for, even if they don't really know what they are going to get. The same goes for becoming a doctor, learning to play the violin, deciding to receive a cochlear implant, and many other aspirations that interact with our core commitments. Another difference between transformative experience and expression concerns the fact that transformative expressions might not be *epistemically* transformative experiences, when they do not involve a new phenomenal content. Nothing in conditions (i)-(3) require that new phenomenal contents play a role, though they often do. However, as we will see in more detail in the following section, transformative expressions change either the content or order of our core commitments and so are personally transformative experiences.

3. Participatory Art

We can broaden the notion of transformative expression to one of transformative *action* by modifying (2) and replacing "expresses" with "is endorsed by." So what does expression add to action in the definition? I might really hate doing the dishes, but after doing them time and again and further appreciating the result of a clean kitchen, doing the dishes might be accommodated by my core commitments. But doing the dishes is still not expressive of who I am as an individual. Various job requirements, duties, practical necessities, personal accommodations, and the like have this feature. Transformative expression has a particular value that transformative action lacks. The idea of a transformative *expression* is that the action expresses some feature of a self that the person did not have before performing it—and that by or through acting, the person comes to realize that the action or something it expresses or embodies is so significant to her as to be self- or life-affirming.

It is this value, I want to argue, that is sought out in a significant tradition of avant-garde art. Much of the rhetoric of the twentieth-century avant-garde concerns transformation. Artists sought to change "subjects," transform "perceptions" or "consciousness," change lives, or transform or "sculpt" society. And to do this they focused much of their creative effort on direct engagement with individuals, constructing "situations," social "interventions," or "happenings" that encouraged the audience to transformatively express themselves: Dada, Futurism, Surrealism, Situationist International, Happenings, Joseph Beuys's "social sculpture," and the more recent work in "relational aesthetics," "social practice," and "participatory art."

Consider this passage from Situationist International co-founder Guy Debord:

> The really experimental direction of situationist activity consists in setting up, on the basis of more or less clearly recognized desires, a temporary field of activity favorable to these desires. This alone can lead to the further clarification of these simple basic desires, and to the confused emergence of new desires whose material roots will be precisely the new reality engendered by situationist constructions. (2006b [1958]: 49)

Debord defined a "situation" as a "concrete construction of momentary ambiances of life," sites of often playful interaction that spoke to "recognized desires." But these situations were designed to appeal to and unmask those desires in order to ground "new desires" that only a transformed society or "new reality" would satisfy—thus creating a collective desire for this new society.

To this end the situationists envisioned various activities meant to interfere with a person's routines and habits and encourage them to embrace a more playful and passionate way of life. "The goal of the situationists is immediate participation in a passionate abundance of life by means of deliberately arranged variations of ephemeral moments" (Debord 2006c [1958]: 53). The situationists envisioned bringing people together and intervening in the gathering to create a more free, playful, game-like atmosphere. They hoped that doing so would allow us to find ourselves, to construct what the situationists called "real individuals" (Debord 2006b [1958]: p. 51) and to focus our creative and playful activity on our own authenticity and community rather than on a material culture that merely entertains, alienates, and divides. "The point is to produce ourselves rather than things that enslave us" (Debord 2006c [1958]: 53).

The situationists are explicit about the interpersonal character of "producing ourselves"; it requires aesthetic attention to shared space and directed intervention:

> If we imagine a particular situation project in which, for example, a research team has arranged an emotionally moving gathering of a few people for an evening, we would no doubt have to distinguish: a director or producer responsible for coordinating the basic elements necessary for the construction of the décor and for working out certain interventions in the events.
>
> (Debord 2006b [1958]: 50)

For the situationists, self-production is a kind of co-production because selves must be playful, expressive, and open, and they conceive of such action as paradigmatically communal.

We can make some of these ideas more concrete by looking at how social intervention and "producing ourselves" is present in the work of conceptual artist and philosopher Adrian Piper. Piper is a light-skinned black woman who is often assumed to be white. As a result, she frequently found herself among white people who thought she was "one

of us" and so would be receptive to their racist conversations and attempts at humor. When she would respond by explicitly calling out the racist or alerting people to her racial identity in advance, she would be perceived as "pushy, manipulative, or socially inappropriate." So she developed an alternative response by intervening in the social dynamics in a way that invites a transformative expression. When someone engaged with, or in, racist talk she would hand out a card that reads:

> Dear Friend,
>
> I am black.
>
> I am sure you did not realize this when you made/laughed at/agreed with that racist remark. In the past, I have attempted to alert white people to my racial identity in advance. Unfortunately, this invariably causes them to react to me as pushy, manipulative, or socially inappropriate. Therefore, my policy is to assume that white people do not make these remarks, even when they believe there are no black people present, and to distribute this card when they do.
>
> I regret any discomfort my presence is causing you, just as I am sure you regret the discomfort your racism is causing me.
>
> Sincerely Yours,
>
> Adrian Margaret Smith Piper[7]

Piper's title for this work—which she calls "reactive guerrilla performance"[8]—is "My Calling (Card) #1." Putting the word "Card" in parentheses draws attention to the fact that this is a calling, and Piper nicely draws on the ambiguity of that word. To "call" on someone is ambiguous between a moral demand and a hopeful plea. Piper's social interventions are more than a calling out; they are a calling for. They shift the social dynamics and prime, indeed invite or call on, the recipients to act in ways that could be transformatively expressive. Receiving a personal and unexpected note is intriguing, mysterious. Piper's design ensures that the first thing one notices after receiving a "gift" is that one is being addressed as a "friend," but one who is invited to reflect on their actions. Piper's intervention gives the friend an opportunity to disavow or distance themselves from what they have done.[9]

This nicely fits the scheme of transformative expression. Prior to Piper's intervention, laughing at racist jokes is compatible with their identity. Maybe they don't have a commitment to being openly racist, but they might have a commitment to expressing solidarity with people of their own race through humor, or perhaps to using collective laughter to reinforce social bonds, even at the expense of marginalized groups. So disavowing racist jokes is not endorsed by their core commitments. If Piper's intervention

[7] For a nuanced discussion of this work, see Marriott (2013).
[8] See Piper (1999: 219).
[9] My technical term for such interventions is "social opening."

is successful and they disavow their actions, then they are in a position to change their core commitments. For some participants, once might be enough; for others, a little time and repetition will make racist jokes incompatible with their core commitments. And this is a direct result of their disavowal in response to Piper's social intervention.

Art historian Grant Kester writes of Piper's work:

> When we encounter new experiences we undergo a transformation, only to gradually recohere around this transformed identity in anticipation of encounters yet to come. The extent to which we are willing to allow these experiences to touch us and to reconfigure our subsequent interactions with others varies from person to person. Piper's performances and installations provide a mise-en-scene designed to encourage such transformations. (2013: 77)

"My Calling (Card) #1" is designed to elicit the knee-jerk defensive rationalizations and stop them in their tracks via the presence of Piper herself—the one who, in calling on this person and creating a social opening, made herself present as a black woman: "My purpose is to transform the viewer psychologically, by presenting him or her with an unavoidable concrete reality that cuts through the defensive rationalizations by which we insulate ourselves against the facts of our political responsibility" (Piper 1999: 234).

Another artist whose work focuses on social and personal transformation is Stephen Willats.[10] In a work entitled *Brentford Towers,* Willats collaborated with residents of a West London apartment complex composed of large uniform beige towers described as "monumental objects" that "seemed to deny the complexity of people's lives within it" (Willats and Ginsborg 2008). Willats thought that this affected the residents' sense of community by affecting their visibility to one another as individuals rather than as little more than "tower resident." Much of Willats' work seeks to engage people in ways that construct or elicit individuality in a way that promotes community: "My work engages the audience in a new way of encountering art in society. I am not talking about a compliance, but something more active, a mutual understanding, an interaction between people—similar to the dynamic image of the homeostat where all the parts of the network are equal and equally linked."

Willats met with fifteen of the individual residents to discuss their living spaces. He asked them to choose a meaningful object from their homes and discussed the importance of this object to them. They then discussed how it related to something outside of the residential tower that the resident could see from their balcony or window. Willats used this information to create a visual artwork that displayed images of the resident, the tower, the interior object, and the exterior object, with lines connecting each and a quote from the resident about his or her life at Brentford Towers: "I need to be out there sometimes. It gives me a taste of what I need, just to get in contact with

[10] For discussion of Willats and awesomeness, see Riggle (2017b: 188-9).

the elements. I like the wind blowing in my face. It makes me feel so much freer." Every two days over a month a new work was displayed on a new floor of the tower, starting from the ground and moving up until they reached the top. The result is what Willats describes as a "sculpture in time moving up through the tower," breaking out spire-like through the top of the tower while representationally (through the displayed works) connected to cultural life outside. This created a new tower, in this case one that was "based on the personal conceptualizations of the tenants, of their daily lives within the building rather than the conceptualizations of architects and planners." This in turn created a symbolic sense of collective ownership of this new "conceptual" tower, a sense of collective ownership of space meant to enrich the sense of community among the residents.

Willats draws attention to the difficulties of "presenting oneself" in certain conditions, and uses various methods of interpersonal connection—photography, discussion, use of space, meaningful attachment, public display—to spur and further such expression. Again this fits the scheme of transformative expression. Roughly, prior to engaging with Willats's work, the residents of Brentford Towers were disinclined to engage with the tower community. After engaging with Willats's work, engaging with the community was among the things they cared about. And this new preference was a result of expressing themselves as individuals-in-a-community through Willats's work.

Some artists focus on individual transformative expression, or artworks as social interventions that target individuals as such, though often with larger social transformation in mind. But creative acts that aim at transformative expression can directly target person-types and groups—and doing so can institute social change.

Consider the norms governing a "person in public" or governing what it is to be a "citizen" in a certain nation. Such a person is governed largely by what Iris Murdoch calls "ordinary public reasons" (1970: 41). They do what "a member of the public" does. The traits and skills one needs to be a person-in-public are neither individualizing nor (normally) partly constitutive of our individual core commitments. Furthermore, the norms that determine what it is to be a person in public are standardly underwritten by broader social, cultural, and political structures. One tactic of social change focuses on getting the "person-in-public" or the "citizen" or other such person-type to transformatively express herself as that type, so as to change the character of the type itself, and to thereby change the character of citizenship.

For an illustration of this, consider Antanas Mockus, who was a math and philosophy professor, served as president of the National University of Colombia, held two terms as mayor of Bogota (1995-7, 2001-3), and was nearly the president of Colombia. Mockus became the mayor of Bogota in 1995, a time when the city suffered enormously. Pedestrian deaths soared, caused by chaotic traffic and little respect for the rules of the road. Violent late-night fights broke out regularly, accompanied by high rates of homicide. There was corruption at every level of governance.[11]

[11] For discussion of Mockus in the context of being 'awesome' see Riggle (2017b: 56-9).

The public citizen of Bogota circa 1993, or at least one widely accepted type of public citizen then, preferred not to observe municipal laws or norms of public respect and order, and generally looked out for himself. This was true of people all throughout all levels of the city's organization, from everyday citizens, to traffic police, businesspeople, and lawmakers.

Bogotá was in need of serious change, but no one knew what to do, and nothing seemed to work. Mockus was especially willing to try anything. He also had almost no experience in politics, and was known for his brazen and unusual leadership of the National University.[12] One of his first projects as mayor was to don a superhero costume with a large yellow C on the chest, which stood for "Super Citizen." A film crew followed him around as he roamed the streets picking up garbage. Mockus also issued 350,000 colorful thumbs-up and thumbs-down cards to be used by drivers. Those breaking or observing the traffic rules could face scores of thumbs-down or thumbs-up signs popping out of car windows.

Mockus suspected that Bogotáns would be more responsive to social stigma and collective action than to tickets and fines. The traffic police were deeply corrupt, which was a major cause of so much disorder in the city. Mockus fired them and replaced them with thousands of mimes. The mimes ran around the city mocking people who violated traffic rules, littered, fought, jaywalked, and so on. When someone did the right thing—crossed in the crosswalk, stopped at the light, threw garbage in the garbage can—the mimes banded together to create celebratory mini-parades and scenes of spontaneous joy. Citizens became witness and participant to these scenes of mockery and praise, joining in on and encouraging the civic fun. Mockus' experiments in transformative civic expression worked: traffic deaths dropped significantly, drivers began to observe the traffic rules, and late- night violence declined (among other things). Mockus had the creative insight to construct situations where people were encouraged to express themselves as citizens in a new way—one that could (and would) transform what it is to be a citizen in Bogotá.[13]

These examples show that participatory art calls on us to be playful, self-reflective, exploratory or adventurous, spontaneous, open-minded, to engage in make-believe, be creative, imaginative, or engage in exploratory dialogue with strangers (among other things). What these activities have in common is that they bear a certain relation to our core commitments: they distance us from our core commitments in a way that allows for transformative expression.[14]

So what makes an artwork transformatively expressive? One option is to say that a work is transformatively expressive iff it causes transformative expressions in participants. However, it is too demanding of these works to require success in changing

[12] He resigned in 1993 after he mooned a group of rioting students who refused to listen to administration leaders at an assembly. He promptly ran for mayor.

[13] Bogota's famous "comeback" is still precarious. In 2011 the *New York Times* reported a return of corruption and traffic chaos. See Romero (2011).

[14] Some philosophers have made use of the "distancing" feature of make-believe. For an account of

people's core commitments. A nice chat with a few strangers in a museum might affect my core commitments—get me to question them, reflect upon them, imagine having different ones or restructuring them—but it probably won't change them permanently. I would have to remain inspired by the work and act on that inspiration. And whether I can depends on so many factors—social, political, economic, personal. As a result, the causal proposal does not adequately capture failures of uptake. Suppose Piper's work never actually transforms anyone because her participants turn out to be too defensive, stubborn, or mean. In that case the causal model would count Piper's work as a failure insofar as she aimed to create a transforma- tively expressive work. But that's the wrong result. The failure is due to the participants, not Piper.

A better option is to say that a work is transformative iff it invites transformative expression. How can a work do that? A work can invite transformative expression by inviting a certain kind of uptake in participants, namely, engagement with the work through the kinds of actions the work requires. On this model, Piper's "guerrilla performance" is a transformatively expressive work in virtue of its inviting the kind of engagement that transformative expression requires. Insofar as Piper's participants fail to take up her artistic invitations, any fault lies with them. So a work is transformatively expressive in virtue of the kind of activity it invites—the kind that is unified by its distancing relation to our core commitments—whether or not it actually transforms anyone. For such a work to fail as a work of transformative expression, it must fail to invite the kind of activity it is designed to invite.

As our examples show, not all works are transformatively expressive in the same way. We can categorize different kinds of transformatively expressive artworks according to the kind of transformation the work aims to foster.

> *Replacement*: The work is designed to replace at least one core commitment with another. Debord's artistic and political aspirations often suggest this approach.
>
> *Introduction:* The work is designed to introduce a new core commitment. Willats's works, especially *Brentford Towers*, aims to cultivate a new commitment for the community-building character of individual expression.
>
> *Elimination:* The work is designed to eliminate a core commitment. Piper's "My Calling (Card) #1" might introduce a new commitment, but it could succeed without doing so. A more fundamental aim is to rid people of the comfort they feel with racist comradery.
>
> *Structure*: The work is designed to restructure commitments. This can happen in two ways:
>
> *Core Restructuring*: Core commitments are reordered in a transformative way.

Peripheral Restructuring: Peripheral commitments are made core. Whether core structural change is, strictly speaking, transformative expression depends on how the pre-transformation commitments fail to endorse the actions endorsed post-transformation. Suppose I have a core preference ranking of food: (1) red meat, (2) poultry, (3) fish, (4) grains, (5) vegetables. On this ranking, I genuinely love vegetables—I don't have a mere peripheral preference for them. However, whenever meat is available—red, poultry, or fish—I will choose it over grains and vegetables. There's a holistic sense in which my choosing vegetables is not endorsed by my core preferences: When all my preferences are taken into account, eating vegetables is generally not what I will do. But suppose one night a brilliant friend cooks me the best vegetarian meal of my life—one that changes my sense of possibility for eating vegetarian. It's not enough to convert me; I retain all of my core food preferences. However, the meal succeeds in restructuring them: Vegetables and grains now come first and second respectively, and I almost always choose them over meat. This restructuring is likely to result in robustly different patterns of action when it comes to food.

Group Transformation: The work is designed to cause a group of people to share commitments, where the pre-transformed group exhibits a plurality of commitments and is transformed into a homogeneity. This can be done in positive (Mockus) and negative ways through various combinations of the above.

Of course, some works are more complex or dynamic and involve multiple kinds of transformative expression. For "Documenta 12," Chinese artist Ai Weiwei created "Fairytale" (2007). He flew 1,001 Chinese citizens to Kassel to spend time in a European city, explore the art, and interact with German residents. His goal, as stated in an interview about the work, was "[t]o let them look at each other; to let them have an imagination about each other; to have romance and fantasies about each other" (dmovies.net 2013). A work this open-ended can be transformatively expressive in any of the ways detailed here.

Although it is illuminating to understand this avant-garde tradition in terms of transformative expression, it would be a mistake to think that this exhausts its interest and value. Despite the avant-garde's emphasis on transformation, these works often do far more than that: and those that are not transformative might succeed in other ways. In this, transformatively expressive art has much in common with religious and public art that uses aesthetic techniques and designs to attract, transform, and commune. As Western institutions secularized and museum and gallery practices ascended to promote and support the "fine arts," these expressive practices emerged in other ways

action by ideal that does so, see Velleman (2006). For a discussion of, and alternative to, this account, see Riggle (2017a).

with things like social protest music, memorial art, and mural and street art—practices that flourished outside of artworld confines.[15] One way to think about participatory art is as the artworld-sanctioned secularization and pluralization of expressive community building.

4. Aesthetic Value and Action

By the late twentieth century, the philosophical and artworld understanding of art had shifted so far away from its expressive, communal role that French curator and writer Nicolas Bourriaud, an early theorist of and advocate for the participatory or (as he called it) "relational" shift in art, could write:

> Today, [art] history seems to have taken a new turn ... artistic practice is now focused upon the sphere of inter-human relations, as illustrated by artistic activities that have been in progress since the early 1990s. So the artist sets his sights more and more clearly on the relations that his work will create among his public, and on the invention of models of sociability ... Meetings, encounters, events, various types of collaboration between people, games, festivals, and places of conviviality, in a word all manner of encounter and relational invention thus represent, today, aesthetic objects likely to be looked at as such ... "
>
> (1998: 28-9)

But what conceptual resources do we have for thinking that the works considered here have aesthetic value? They don't fit common ways of thinking about art and aesthetic value: They involve social interactions and processes, many of which do not or even could not exist in a museum or gallery. Many of the best works are not visually or aurally pleasing, if only because there is little to contemplatively view or hear. And relatedly, their effects are often not traditional aesthetic excitements, thrills, or the sense of calm or wonder familiar from many good artworks. These works focus on transformatively expressive actions with the hope of deepening connections, enriching community, and cultivating individuality, mutual understanding, and interpersonal appreciation.

To be clear, this is a question about the aesthetic value of these works, not artistic value. Artistic value is one thing, aesthetic value another. Lots of non-art can possess aesthetic value: sunsets, people, flowers, natural sounds, landscapes, animals, waves, skies, succulents, stars, rocks, seashells, coral, etc. Some artworks have pro tanto aesthetic merit but lack all-things-considered artistic value: for example, some film critics argue that Terrence Mallick's later films are beautiful but defective films. Furthermore, and this is the main point, something's *artistic* value might have nothing to do with

[15] See Wolterstorff (2015); also see Riggle (2010).

its *aesthetic* value, whether or not it possesses such: John Cage's 4'33", Duchamp's *Fountain,* Walter de Maria's "Vertical Earth Kilometer."[16]

I do not doubt that the socially engaged works considered here have *artistic* value—they are good artworks.[17] The difficulty in properly understanding these works concerns the thought that they have *aesthetic* value. Indeed, they are not just artworks with social or ethical intent and effect; some of the best of them are *beautiful.* They are "aesthetic objects likely to be looked at as such." The Situationists explicitly sought this "new beauty": "The new beauty will be *the situation,* that is, temporary and lived."[18]

So here's the difficulty: how can we say that these are aesthetically good, even beautiful artworks when their value lies primarily in their transformatively expressive character? That might confer *artistic* value, but how might it confer aesthetic value? How can we capture this "new beauty"?

It's not obvious, and traditional ways of thinking about aesthetic value are of little help. When we ask what aesthetic value is, there are two questions we might be asking. If we are asking what makes gracefulness, elegance, sleekness, or smoothness aesthetic values, we might be granting that they are values and asking what qualifies each as *aesthetic* (the demarcation question), or we might be granting that they are *aesthetic* values and asking what their value consists in (the normative question).[19]

A traditional response to the demarcation question is "formalism," or the view that strictly aesthetic value lies in certain properties that supervene on formal or "configurational" properties: The elegance of a sculpture supervenes on its shape; the warmth and softness of a painting supervenes on the paint hues and the lack of definition in the lines. The traditional response to the normative question is "hedonism," or the view that aesthetic value bears a constitutive relation to pleasure. But what are the aesthetically relevant configurational properties of, for example, Piper's "My Calling (Card) #1"? There are some, to be sure—e.g. the aesthetic properties of the card that Piper designed and passed out—but they do not capture the aesthetic value of the performance. Nor does a focus on pleasure seem apt here. Piper's actions are occasioned by social injustice and they are intended to confront, challenge, and transform.

It is no surprise, then, that there is a strong inclination to think about the value of these works in ethical terms. Grant Kester understands them in terms of the "creative

[16] At least, the burden of proof is on those who deny this claim. For support of the view that artistic value is a kind of aesthetic value, see Shelley (2003) and Lopes (2011). For responses to Lopes, see Hanson (2013) and Huddleston (2012).

[17] For a take on what their artistic value consists in, see Simoniti (2018).

[18] "La poesie a epuise ses derniers prestiges formels. Au-dela de l'esthetique, elle est toute dans le pouvoir des hommes sur leurs aventures. La poésie se lit sur les visages. Il est donc urgent de créer des visages nouveaux. La poésie est dans la forme des villes. Nous allons donc en construire de bouleversantes. La beaute nouvelle sera DE SITUATION, c'est-a-dire *provisoire* et vecue." "Reponse a une enquete du groupe surrealiste belge" ("Quel sens donnez-vous au mot 'poesie'?," 1954). This is from the Letterist International, the ideologically similar group that immediately preceded the Situationists.

[19] See Lopes (2018: 41-3).

orchestration of dialogical exchange" (2013: 189), where dialogical exchange "requires that we strive to acknowledge the specific identity of our interlocutors and conceive of them not simply as subjects on whose behalf we might act but as co-participants in the transformation of both self and society" (2013: 79). Kester's discussion and analysis of these socially oriented artworks is extremely valuable, but the same question arises: What exactly is "aesthetic" about the "creative orchestration" of dialogical exchange? And how do we unpack that metaphor? Interpersonal recognition for the sake of positive personal and social transformation is something whose value we should understand in ethical and political terms. Kester notes the strain on our traditional understanding of art and the aesthetic: "There is potentially productive terrain here for an expanded analysis of the aesthetic" (2013: 189). But, again, what could that "expanded analysis" be?

Art historian and critic Claire Bishop makes a similar point when she notes that these works don't sit comfortably in either category of the aesthetic or the ethical:

> contemporary art's "social turn" not only designates an orientation towards concrete goals in art, but also the critical perception that these are more substantial, "real", and important than artistic experiences. At the same time, these perceived social achievements are never compared with actual (and innovative) social projects taking place outside the realm of art.
>
> (Bishop 2012: 19)

Critics who set their sights on such works tend to use sociological and ethical categories like empathy, identification, community, social and personal transformation. But they do so while insisting that these works are art and so ought to be evaluated as such rather than as social or political initiatives intended to bolster empathy, community, social transformation, and so on. If we focus on what seems to make these works important—their "concrete goals" for personal and social change—then why not critically compare them with innovative social projects in general? With social or municipal programs, non-profit initiatives, or legislation?

Kester's and Bishop's remarks suggest the challenge: we need to find a way of thinking about aesthetic value that (1) makes sense of the aesthetic value of transformatively expressive works; (2) does so in a way that captures the thought that the aesthetic character of these works is ethically significant; yet (3) justifies the critical impulse to compare these works to other items of aesthetic value.[20]

A move in the right direction comes from Sarah Hegenbart, who proposes a "virtue account" of participatory art: "Whereas Socrates elicits a refinement of the virtues

[20] Bishop addresses the issue, but her suggestive proposal is underdeveloped. She rejects understanding such works in ethical or political terms on the grounds that doing so "fails to accommodate the aesthetic or to understand it as an autonomous realm of experience" (2012: 40). She associates a concern with ethics and politics with terms like "social obligation," "super ego," "guilt," "self-suppression," and "social consensus." And she contrasts this with the thought that we find a certain joy in acting on

through the rational activity of critical thinking, the [participatory] artist prepares a platform for the refinement of the virtues through aesthetic engagement" (Hegenbart 2016). What Hegenbart has in mind with "aesthetic engagement" are the creative and imaginative skills required to constitute the "meaning" of the artwork as intended by the artist (pp. 333-4). To make this concrete, consider *Brentford Towers.* Willats's intention was, in large part, to create a "conceptual" tower through the expressive participation of, and collaboration with, the tower residents. To participate in the creation and continued existence of this tower, the residents had to open themselves up to Willats and their fellow tower residents and contemplate and express their individuality, reflecting on and explaining décor choices, considering their visual perspective from the tower on the outside world. They also had to exercise their appreciative skills when the diagrams were displayed, interact with one another about the creation and presence of the work, and so on.

Hegenbart proposes that we understand the aesthetic value of such works by connecting these creative and expressive activities to virtue. She adopts a welcoming theory of virtue according to which virtues are "excellent skills that enable us to overcome ... novel challenges."[21] And using this theory, she puts the aesthetic value of participatory works down to a connection between ethical and aesthetic virtue: "If we practice creative responses to new and unexpected situations in the aesthetic realm, this might improve our ability to respond creatively to new situations in the moral realm. So creativity may enable us to react reliably across different moral and aesthetic situations" (2016: 335).

The idea is that participatory works have aesthetic value because they cultivate aesthetic virtues. And aesthetic virtues, in turn, are important for the cultivation and practice of moral virtue. But does this answer the challenge? It seems not. While Hegenbart offers a way of making sense of (1) and (2), her view does not seem to justify the critical impulse to compare these works to other works of aesthetic value. Any project aimed at cultivating aesthetic virtues—art classes, dance contests, after-school arts programs, and indeed style and cooking shows—would seem to make fair comparisons.

While I don't doubt that participatory works could make participants more virtuous in some sense, there is a deeper problem with understanding the distinctive value of participatory works in such terms. If we employ a moral concept of virtue, then the theory will be too narrow to capture the range of ways these works engage participants

uninhibited desire. She favors thinking about participatory works (the good ones at least) as works that do not shape or suppress desire but rather liberate it, or accommodate it in raw form, and thereby offer such pleasure. Bishop seems to think that there is personal and communal value in works that function this way, but she it is not clear about what the value amounts to. And by emphasizing individual "enjoyment" and the "autonomy" of the aesthetic, she seems to rely on hedonism and formalism (or at least on thoughts that tend to motivate those views) and to reject Kester's call for an "expanded analysis of the aesthetic." As a result, it is not clear how her view meets the challenge.

[21] She takes inspiration from Julie Annas' account of virtue developed in Annas (2011).

in ethically salient activities. As we noted above, many participatory works are intent on encouraging, among other things, play, expressive freedom, adventurousness, and self-exploration through novel, challenging, or creative action. These activities implicate us in interpersonal relations that can ground community and that are normatively structured in substantive ways, but which escape traditional ways of thinking about morality. But if we employ a broader concept of virtue—as something like human excellence—then we will misconstrue the value of these works. They are valuable, ethically and aesthetically, independently of any connection they might have to the good of certain cultivated dispositions that (partly) constitute moral excellence. Indeed, our discussion thus far shows that part of their value lies precisely in distancing us from, or breaking us out of, our dispositions.

Let's make this more concrete by focusing on one of the relevant action-types— play. Play is good, and part of its value involves the kind of volitional openness that lends itself to transformative expression. The ability to play is central to the cultivation, expression, and mutual appreciation of individuality; it allows us to explore different ways of cultivating ourselves—our values, ideals, projects, and so on; and it allows us to creatively riff on, or break out of, social norms and everyday routines to express our individuality to others, so that we can connect with one another as individuals, not merely as people playing social roles or as subjects who merit respect. Being playful is part of what it is to be an individual.[22]

But we can be playful without playfulness featuring among our core commitments. The ability to play does not require that we be a playful person. The capacity to play is a kind of ever-present foil against or alongside our core commitments. In other words, to be playful is not necessarily to have a commitment or disposition to play. It's to have an ability to freely disengage from your commitments, dispositions, preferences, and so on.

To illustrate this further, let's return to style-makeover Dagny. Imagine that Dagny is motivated in part by a playful spirit. We don't have to think that Dagny becomes a more virtuous person in order to think that her playfully adopting a new style is ethically and aesthetically significant. Cultivating and expressing our individuality plays an important role in interpersonal relations. Dagny's new style will create social openings for those who are acquainted with her; inspire others in their own style activity; create new opportunities for gifting, exploration, and exchange. Her new way of expressing her individuality will be a focus of her own creative efforts, seeping into a number of other ways she has of expressing herself, but in a way that binds her to others through shared expressive and appreciative activity. And all of these activities are subject to interpersonal normative structure that determines the appropriateness of various modes of response to Dagny and her new style. Focusing on the question of whether Dagny's playful change makes her a more excellent person obscures the

[22] For discussion of this in the context of a theory of social virtue, see Riggle (2017b).

personal and communal import of her efforts, and threatens to shift focus away from the tentative, experimental, exploratory character of these activities.

When we play, we engage what I call our "pure individuality," which forms the core of what I think of more generally as our individuality. Our pure individuality is our basic capacity to act in a volitionally open way—to imitate, be spontaneous, adventurous, expressive, and so on. These activities allow us to reflect on, refine, and cultivate our core commitments; and they help us change or abandon them. Our individuality gains definition as we cultivate ways of playing, exploring, etc. that we value—we cultivate not just the basic ability to value and appreciate, but a refined sense of care and love; not just the basic ability to spontaneously imagine, but the capacity for make-believe and storytelling; not just the basic ability to joke and laugh, but a sense of humor; and so on. In other words, we use our pure individuality to sculpt a self. It is through these characteristically aesthetic activities—play, adventure, experimentation, spontaneity, and so on—that we shape, cultivate, and reshape ourselves as individuals. In this way, acting from our pure individuality, and cultivating the ability to do so, is ethically significant.

So in what sense, if at all, do these actions have aesthetic value? One approach is to try to tie them to traditional ways of thinking about aesthetic value. We already saw that hedonism and formalism have difficulty with transformatively expressive works. But another thought behind such theories is that it is only when we experience configurational properties in a certain way, "disinterestedly," that we can experience the aesthetic properties that supervene on them and so properly experience the pleasures they occasion. Disinterest is the glue that binds formalism and hedonism: To experience the grace in a speech, the elegance of a sculpture's shape, the sleekness of a coat, one must experience the speech, the sculpture, the coat in such a way that one's various motives and commitments do not get in the way. For example, if you disagree with the content of the speech, are too focused on how the sculpture will make you money or will socially impress, or find umbrellas useless because you live in a desert, your focus on these features threatens to prevent you from taking pleasure in the grace, elegance, or sleekness. Following this way of thinking, we might say that it is only when we take up the invitations of transformatively expressive art—distancing ourselves from or bracketing our core commitments, i.e. "disinterestedly" as it were—that are we in a position to be affected by those works as intended. Where certain properties of works invite us to see or hear or more generally experience in certain aesthetic ways, so other properties invite us to act in certain aesthetic ways. But in both cases a kind of "disinterest" is important. This could justify the critical impulse to place participatory art in an evaluative comparison class with other works of aesthetic value.

Another approach abandons these ways of thinking about aesthetic value.[23] Instead of thinking of aesthetic value as the power to please, we might think of it as the

[23] For a critique of disinterest, see Riggle (2016); for a critique of hedonism, see Lopes (2018); see also Van der Berg (forthcoming).

power to put us in a more volitionally open state. On this view, the aesthetic value of transformatively expressive art lies in its engaging us in distancing activities, inviting us to freely disengage from our core commitments. What makes these works aesthetically valuable is not that they have certain configurational properties on which aesthetic properties supervene; rather, it's that they invite aesthetic action. Where some works invite us to see, hear, or experience certain things, others invite us to do certain things. And when what they invite us to do has aesthetic value, we can say that such works are good artworks and are good in virtue of their aesthetic value.

5. Conclusions

Thinking about aesthetic value and action in this way helps us answer the the question about action that we posed at the outset: If a transformatively expressive action is not endorsed by, and is in some cases incompatible with, one's core commitments, then how can we say that it is self-expressive? In what sense is a "self" being expressed, even though we aren't expressing our core commitments? We might even worry that we cannot say that the action is one's own or a product of selfgovernment and not alienated, forced, or just something that happens or is caused by the artwork.

Our discussion suggests that there's a part of the self we can tap into even when (but not always when) we have lost touch with our core commitments. In freely distancing ourselves from our commitments, we tap into another part of ourselves. When we express our pure individuality, we express deep features of ourselves in ways that might fill gaps in, or run counter to, our cultivated sense of self but precisely because of that might be liberating, even transformative. We might come to see a new side of ourselves or of life. This amounts to a counterexample to various partial theories of authenticity and autonomy that require the expression of self-constituting dispositions. If what I have suggested is right, then that part of the self is aesthetic. If we want to understand human action, then we cannot ignore the aesthetic.

Of course, there is much more to explore and pin down along these lines. But the thought that there is a deep connection between aesthetic value, self-realization, and volitionally open action is not without precedent. Schiller thought that really great art would have this kind of distancing effect on us: through witnessing genuine art we enter a volitional state in which we "shall with equal ease turn to seriousness or to play, to repose or to movement, to compliance or to resistance." Schiller thought that doing so makes us more social and egalitarian individuals, equipped to engage in activities that allow us to present, recognize, and appreciate our own and each other's individuality. Schiller called this special state "play."

Schiller also tells us, "A person only plays when they are a person in the full sense of the word, and they are fully a person only when they play." As strange as that may sound, maybe there's something to it. There is a tight connection between the pure individuality and a capacity that has as good a claim as anything else to being that

which grounds our personhood, namely, our capacity for love. If we are so attached to our core commitments that we lose touch with our pure individuality—unable to play, pretend, be spontaneous, adventurous, and so on—then we also lose access to the forms of interaction and mutual appreciation that depend on it. The early Situationists emphasized this connection: "[T]his striving for playful creativity must be extended to all known forms of human relationships, so as to influence, for example, the historical evolution of sentiments like friendship and love" (Debord 2006a [1957]: 40). Schiller himself thinks that being in touch with aesthetic value, i.e. that which causes us to play, is the only way we can really be free, and that this freedom is what makes us truly social. That we all attain such a state is a hope embodied in so many transformatively expressive works—the hope that, as we cultivate ourselves and each other through aesthetic acts and objects, we will find new and exciting ways of being and bonding.[24,25]

References

Annas, J. 2011. *Intelligent Virtue.* Oxford: Oxford University Press.

Bishop, C. 2012. *Artificial Hells.* London: Verso.

Bourriaud, N. 1998. *Relational Aesthetics.* Dijon: Les presses du reel.

Calhoun, C. 2009. "What Good Is Commitment?" *Ethics* 119(4): 613-41.

Carter, R. 2008. *The Japanese Arts and Self-Cultivation.* New York: SUNY Press.

Chung, J. (2018). "Moral Cultivation: Japanese Gardens, Personal Ideals, and Ecological Citizenship." *Journal of Aesthetics and Art Criticism* 76(4): 507-18.

Clark, T., C. Gray, D. Nicholson-Smith, and C. Radcliffe. n.d. "The Revolution of Modern Art and the Modern Art of Revolution." <https://www.cddc.vt.edu/sionline/si/modernar

Debord, G. 2006a [1957]. "Report on the Construction of Situations and on the International Situationist Tendency's Conditions of Organization and Action." In K. Knabb (ed. and trans.), *International Situationalist Anthology,* 25-43. Berkeley, CA: Bureau of Public Secrets.

Debord, G. 2006b [1958]. "Preliminary Problems in Constructing a Situation." In K. Knabb (ed. and trans.), *International Situationalist Anthology,* 49-51). Berkeley, CA: Bureau of Public Secrets.

Debord, G. 2006c [1958]. "Theses on Cultural Revolution." In K. Knabb (ed. and trans.), *International Situationalist Anthology,* 53-4). Berkeley, CA: Bureau of Public Secrets. dmovies.net. 2013. "Ai Weiwei, Fairytale (2007), interview at Documenta 12." <https://vimeo. com/68202707>

[24] For a detailed interpretation of Schiller's theory of aesthetic value and its connection to individual and political freedom, see Riggle and Matherne (2018).

[25] Thanks to Antonia Peacocke and Eric Carter for helpful comments. Thanks also to audiences at the Chapel Hill Workshop on Transformative Experience and the APA Pre-Conference on Themes in Transformative Experience.

Frankfurt, H. G. 1971. "Freedom of the Will and the Concept of a Person." *Journal of Philosophy* 68(1): 5-20.

Hanson, L. 2013. "The Reality of (Non-Aesthetic) Artistic Value." *Philisophical Quarterly* 63(252): 492-508.

Hegenbart, S. 2016. "The Participatory Art Museum: Approached from a Philosophical Perspective." *Royal Institute of Philosophy Supplement* 79: 319-39.

Huddleston, Andrew. 2012. "In Defense of Artistic Value." *Philosophical Quarterly* 62(249): 705-14.

Kester, G. H. 2013. *Conversation Pieces: Community and Communication in Modern Art.* Berkeley, CA: University of California Press.

Lopes, D. M. 2011. "The Myth Of (Non-Aesthetic) Artistic Value." *Philosophical Quarterly* 61(244): 518-36.

Magritte, R. (ed.) 1954. "Quel sens donnez-vous au mot 'poesie'?" *La carte d'apres nature* (January).

Marriott, D. 2013. "On Racial Etiquette: Adrian Piper's *My Calling (Cards)." Postmodern Culture* 24(1).

Murdoch, I. 1970. *The Sovereignty of Good.* Abingdon: Routledge.

Oshana, M. 2007. "Autonomy and the Question of Authenticity." *Social Theory and Practice* 33(3): 411-429.

Piper, A. 1999. *Out of Order, Out of Sight.* Cambridge, MA: MIT Press.

Riggle, N. 2010. "Street Art: The Transfiguration of the Commonplaces." *Journal of Aesthetics and Art Criticism* 68(3): 243-57.

Riggle, N. 2016. "On the Interest in Beauty and Disinterest." *Philosophers' Imprint* 16(1): 1-14.

Riggle, N. 2017a. "Personal Ideals as Metaphors." *Journal of the American Philosophical Association* 3(3): 265-83.

Riggle, N. 2017b. *On Being Awesome: A Unified Theory of How Not to Suck.* London: Penguin Random House.

Riggle, N., and S. Matherne. 2018. "Schiller on Freedom and Aesthetic Value." Paper presented at Harvard European Philosophy Workshop, December 7, Cambridge, MA.

Romero, S. 2011. "Columbia's Resurgent Capital Backslides Amid Crime and Congestion." *New York Times* (May 5).

Saltz, J. 2010. "How I Made an Artwork Cry." *New York Magazine* (February 7).

Shelley, J. 2003. "The Problem of Non-Perceptual Art." *British Journal of Aesthetics* 43(4): 363-78.

Simoniti, V. 2018. "Assessing Socially Engaged Art." *Journal of Aesthetics and Art Criticism* 76(1), 71-82.

Velleman, J. D. 2006. "Motivation by Ideal." In J. D. Velleman (ed.), *Self to Self*, 312-29. Cambridge: Cambridge University Press.

Walden, K. 2015. "Art and Moral Revolution." *Journal of Aesthetics and Art Criticism* 73(3): 283-95.

Willats, S. (producer), and C. Ginsborg (director) 2008. *A State of Agreement* [motion picture]. England: Arts Council England.
Wolterstorff, N. 2015. *Art Rethought: The Social Practices of Art.* Oxford: Oxford University Press.

10. Learning from Moral Failure[13]

Matthew Cashman and Fiery Cushman

1. Introduction

Military basic training is organized around a simple goal: Within a few months, train an ordinary i8-year-old to follow orders that include risking their own life and ending those of others. To accomplish so much change in so little time would seem laughably ambitious if it didn't work so well.

The recruit must learn new motor skills (e.g. how to fire and maintain a rifle), new semantic knowledge (e.g. rules of engagement), and new social orders (e.g. the chain of command). The most remarkable transformations, however, are moral. Recruits are asked to subordinate personal wellbeing to that of the group; to supplant individuality agency with collective allegiance and hierarchy; to regard harm not as intrinsically wrong but as instrumentally justified. How is this done?

We focus on one small but exquisitely counterintuitive piece of the puzzle: Sometimes, students are set up to fail. That is, instructors occasionally create situations where recruits are unlikely to succeed at a stated objective, and where their failure has moral consequences subject to corrective action by the instructor (Heckathorn 1988). The recruits' failures are often on simple tasks: uniform not in order, bed not made correctly, late to parade. These are tasks that, individually, are well within the capacity of any recruit. Taken collectively—in the context of little food, reduced sleep, copious exercise, and stress—the likelihood of somebody in the squad failing increases. Because these tasks are simple, recruits are prone to attribute failure to some personal defect or deliberate choice; after all, who can't make a bed? Moreover, what would otherwise be non-moral failures become moral failures when punishment is applied to the recruit's entire group. This invites other-oriented emotions such as guilt and shame—a feeling of having "let down my squad."

In other words, basic training seeks to change the moral character of recruits in part by designing situations in which moral failure is likely. An implicit assumption is that the lessons learned from experiencing failure, or witnessing it, are especially powerful

[13] Matthew Cashman and Fiery Cushman, Learning from Moral Failure In: Becoming Someone New: Essays on Transformative Experience, Choice, and Change. Edited by: Enoch Lambert and John Schwenkler, Oxford University Press (2020).

© Matthew Cashman and Fiery Cushman. DOI: 10.1093/oso/9780198823735.003.00011

and enduring. Our goal is to scrutinize this assumption. We ask whether, and when, moral failure can be a productive element of moral education.

In its most essential form, moral failure has two parts: it is a direct experience of subjectively failing to meet moral standards. Forgetting your child's birthday, for instance, would likely be experienced by most people as a moral failure. Of course, much (perhaps most) moral learning does not involve moral failure. First, not all forms of moral learning involve direct experience; an alternative form of moral learning involves exposure to abstract rules (e.g. exposure to the ten commandments). Second, not all forms of learning from direct experience involve failure; an alternative form of learning involves practicing and experiencing success (e.g. cultivating the habit of empathy through meditation).

There are also useful and important ways of relaxing this strict definition of moral failure. If a person reads a novel in which the protagonist experiences moral failure, this may prompt a similar psychological response to personally experiencing moral failure even though it is not "direct experience" of the reader's. Or, if a person performs an action they sincerely believe to be right but notices that everybody else considers it wrong, their subjective experience may be a moral failure of some kind, even if not of the purest kind. We will therefore consider not only prototypical moral failures, but also their many kin.

Here, we build a case for the value of moral failure by drawing on several circumstantial lines of evidence. Direct experience is a privileged form of learning; morality requires representations of what not to do (not just what to do); guilt is an adaptive response to moral failure that facilitates reparation and learning; and child development appears to involve periods designed to "test limits" in a way that will reliably lead to moral failure. Yet there is little direct research on the practical value of moral failure, and several reasons to doubt whether it is especially effective, or whether its obvious costs could appropriately outweigh its potential benefits. We therefore conclude by considering opportunities for further research.

2. Direct Experience Is a Privileged Form of Learning

There are some things you have to try in order to learn. You can't read a book to learn how to water ski, and you can't learn to play a guitar without touching one. Of course, many other things can be learned perfectly well from books, lectures, and the like— things like who won the World Series in 1927. Semantic knowledge of this form rarely depends on experience: To know who won the World Series, it isn't at all important that you attended the game. This reflects the fact that human memory and learning comprise multiple distinct systems. Procedural memories of how to execute the small, finely tuned motions necessary to stay upright on water skis are gleaned

from practice, episodic memories supply the broad strokes (such as how to attach the tow rope and put on the skis), and semantic memory may supply as-yet unused hand-signals (such as for an emergency stop).

At first blush, morality would seem to depend principally upon semantic memory. To understand that it is wrong to kill, isn't it sufficient just to be told so? Surely it isn't necessary to actually kill a person oneself ?

We propose that moral knowledge occupies a point somewhere in between water skiing and knowing who won the World Series: Although having a direct, personal experience of the consequences of an action are certainly not necessary to represent that it is right or wrong, the manner in which you know it and the strength of your conviction may depend upon experience. Perhaps a person who has killed can experience its wrongness in a way that is difficult—even impossible—for others to fully understand.

The key concept that connects procedural learning to moral behavior is value. Several lines of evidence indicate that human moral behavior depends upon representations of value (Crockett 2013; Cushman 2013; Ruff and Fehr 2014). The way that people make moral trade-offs exhibits signatures of general value-guided decision-making mechanisms, such as diminishing marginal returns (the difference between saving zero and one lives feels more profound than the difference between saving 100 and 101 (Shenhav and Greene 2010)). Neural systems implicated in learning and representing value are reliably recruited during moral judgment and decision-making (reviewed in Ruff and Fehr 2014). The disruption of these systems by injury can lead to disorders of moral judgment and behavior (Koenigs et al. 2007; Damasio 1994; Lough et al. 2006; Darby et al. 2017), and the same systems appear to be dysfunctional in psychopaths (Buckholtz et al. 2010; Finger et al. 2008).

Meanwhile, much procedural learning depends upon value representation. Computational models of value-guided learning and decision-making provide an excellent fit to behavioral and neural evidence on forms of procedural learning, including habitual action and thought (Dolan and Dayan 2013; Daw and Shohamy 2008; Glimcher 2011). Experience is the canonical path by which value is learned. Although it is certainly possible to learn value in other ways—for instance, by observing other people, being instructed by other people, or imagining alternative possibilities (Gershman et al. 2014; Olsson and Phelps 2007; Ho et al. 2015; Ho et al. 2016)—direct personal experience appears to be a privileged form of value learning (Paul 2014; Olsson and Phelps 2007). Indeed, there are many domains where learning without experience, as by reading or instruction, will be ineffective. Similarly, there are situations where reading a book or attending class will result in some useful amount of learning—but this is strengthened by practice (Kolb 2014), and many instances where simulations are used in pedagogy to access experiential learning that would otherwise be impractical (Ruben 1999).

Motivated by this pair of observations—that morality depends on value representation, and direct experience is a privileged form of learning value—several recent theories propose that moral behavior is strongly influenced by implicit value representations, often learned through direct experience (e.g. Crockett 2013; Cushman 2013;

Rand et al. 2014). Although recent years have seen renewed interest in this idea, its historical roots lie at least as deep as Aristotle *(Ethics,* 1166b5-29). When applied to the concept of moral failure, the implication is that sometimes *actually* failing will engage an especially powerful form of learning.

3. You Have to Learn the "Don'ts": Proscriptive Morality is Unique

If moral behavior depends upon value representations that can be acquired through experience and practice (something roughly like learning to water ski), then the most obvious implication for moral education would seem to be to practice doing the *right* thing—precisely the opposite of moral failure. Typically, we assume the hard work of acquiring a new skill consists in discovering what to do, rather than discovering what not to do. It seems as if learning what not to do should come for free—just learn to do what's right, and you'll never even think about doing wrong.

Yet, in contrast to this picture, people seem to have dissociable systems for learning what to do and what not to do. These are sometime described as "appetitive" versus "aversive" learning systems, and sometimes in other terms (Carver 2001; Frank et al. 2004). Several lines of evidence indicate the significance of this distinction. They can be broadly, if not perfectly distinguished neuroanatomically. For instance, the striatum plays a more essential role in appetitive responding and the amygdala in aversive responding (LeDoux 2003; but see Seymour etal. 2005), although both structures clearly participate in both, and may be best characterized in terms of different learning rules (Li et al. 2011). Likewise, neurochemically, dopamine seems to play a more essential role in appetitive learning and serotonin in aversive learning (Crockett etal. 2009; Cools etal. 2011; Boureau and Dayan 2011), although again the functions of each neurotransmitter are varied and overlapping. These findings indicate that the human mind does not respond to all motivational influences alike; rather, it respects a rough divide between systems that regulate behavior around "promotion" (i.e. appetite) and "prevention" (i.e. aversion).

This division is also clearly reflected in moral judgment and behavior (Janoff- Bulman et al. 2009; Crockett et al. 2015). That is, people appear to have distinct representations of what is morally required (or at least laudable) and what is morally prohibited (or at least blameworthy). In theory, then, a person could adequately learn to do what is morally good and yet not have learned to avoid what is morally bad.

This implies that even if a person has experienced the rewards of doing the right thing, there is a potential for additional learning by experiencing the consequences of doing wrong. For now, we set aside the vital question of how such pedagogical "value" could ever outweigh the obvious cost of having acted wrongly. Instead, our next goal

is to understand more clearly the specialized psychological mechanisms that help us to learn from moral failure.

4. Guilt Facilitates Learning from Failure

Humans are designed for corrective learning from moral failure, and a linchpin of this design is guilt. Put simply, if you feel guilty about something you did, you're less likely to do it again (Monteith etal. 1993; Mosher 1965; Monteith etal. 2002; Baumeister et al. 1995). This does not imply that guilt is the best path to moral learning, and it is certainly not the only path. But, insofar as guilt is triggered by episodes of self-perceived moral failure, it tends to improve future behavior.

Unfortunately, adaptive moral learning via guilt is not the only possible outcome of perceived moral failure. In opposition to this desired pathway, there is a parallel, undesired pathway from subjective moral failure to social withdrawal and "externalization"—the attribution of blame for the moral failure to external circumstances, rather than the acknowledgment of personal responsibility (Leach and Cidam 2015). (Incidentally, guilt establishes the motive to compensate harms and repair social relationships; although this is not the goal of the desired pathway from moral failure to moral learning, neither is it detrimental to that goal.)

The effort to delineate adaptive and maladaptive responses to moral failure centers on the distinction between "guilt" and "shame." Guilt is typically regarded as a more adaptive response (characterized by learning and reparation), and shame as a less adaptive response (characterized by externalization and withdrawal). We do not mean to imply that either is less biologically or culturally adaptive; likely, both of these mechanisms are "adaptive" in those senses. Rather, we mean adaptive from a social and pedagogical perspective: If you were responsible for somebody's moral education, you would probably wish them to respond to failure with feelings of guilt rather than shame.

It is also important to clarify our use of the very terms "guilt" and "shame," and the status of the distinction we are drawing between them. We do not intend to analyze the folk concept of guilt versus shame, or attempt to understand what these words mean in ordinary usage. In fact, there is very little difference in the way that ordinary people understand or apply the words "guilt" and "shame" (Leach and Cidam 2015). Rather, these have become terms of art in a literature demonstrating that selfperceived moral failures can lead to a series of psychological and behavioral reactions that roughly follow two rival paths. This literature seeks to understand how these paths differ: What determines which path is followed, and where each is likely to lead, both personally and socially. From this perspective, the distinction might as well be described as "Type 1" versus "Type 2" reactions, as opposed to "guilt" and "shame."

Guilt and shame may be widely agreed to be moral emotions and the principal emotions arising from moral failure, but the exact distinction between the two is the subject

of some debate (e.g. Tangney 2002; Tangney etal. 2007; Gilbert etal. 1994; Lindsay-Hartz etal. 1995). There is disagreement about what constitutes guilt vs. shame, their properties, and their relative usefulness. Most commonly, guilt is viewed as the "better" emotion because it results in approach behaviors helpful to the agent and the aggrieved, whereas shame is the "worse" emotion, leading to unhelpful withdrawal behaviors (Nelissen etal. 2013). On some views, guilt and shame are distinguished primarily by their sources (guilt being internal norm violation, shame external), while other, more recent work differentiates based on the object of the emotion (guilt focuses on what has been done, whereas shame focuses on who has done it; Nelissen et al. 2013).

Here, we use "guilt" to describe an emotion elicited by moral failure that focuses on the act rather than the qualities of the person who may have caused it, and which generally results in approach behaviors such as reparation. We use "shame" to describe an emotion that is elicited by moral failures attributed to a defect in the self, which focuses on self-image, and which can lead to either approach or withdrawal behaviors. We therefore treat the emotions that arise from moral failure as organized along a spectrum from those concerned with the act (focused on what has been done: guilt) to those concerned with the agent (focused on who has done it: shame).

How, then, does moral failure lead either to useful motivations (e.g. pro-social or self-improvement motivations) or instead to undesirable ones (e.g. self-defensive motivations)?

Moral failures that are focused on the act rather than the agent ("I can't believe I did that, throwing my recyclables in the trash"), and which are not indicative of some sort of self-defect, can elicit adaptive guilt and therefore prosocial behaviors aimed at self-improvement. Moral failures that are focused on the agent ("I can't believe I was late again, I'm always late") and which are attributed to a specific (likely mutable) self-defect, generally motivate self-improvement via shame. Moral failures that are focused on the agent, and which lead to appraisal of global (seemingly immutable) self-defect ("I caused that crash because I am an alcoholic") engender a type of shame which leads to feelings of inferiority and self-defensive motivations to hide, avoid, and externalize (Gausel and Leach 2011).

Given this framework, we can start to identify a set of moral failures expected to be useful in pedagogy: Those generally resulting in focus on the act itself or those which are attributed to a specific self-defect. Before turning to the practical implications of this view, however, we consider one final piece of evidence that humans learn especially well from the experience of moral failure.

5. Children May Be Designed to Fail

Not only are humans designed to learn from failure, there is good reason to believe that they are actually designed to fail. Colloquially, we speak of children "testing limits." This behavior is especially pronounced during the toddler and preschool years, and then

again during adolescence. To describe child misbehavior as "testing limits" implies that the principal aim of the misbehavior is moral learning. For instance, consider a child who takes a cookie from a jar and begins to eat it in view of her father. Some potential explanations for this behavior are (1) she has not considered the possibility that this behavior is disallowed, or (2) she knows it is disallowed, but doesn't care because she really wants a cookie. But if she is truly testing limits, then an additional explanation is (3) she is unsure whether the behavior is disallowed, or exactly what its consequences will be, and an effective way to learn is to try it. An additional and likely possibility is that (1) or (2) may characterize her proximate psychological motives, while (3) characterizes the ultimate adaptive rationale that explains them.

Of course, children are capable of learning a new rule without first violating it. A child uncertain of the rule could just ask an authority ("Daddy, would it be OK for me to take a cookie right now?," on the questionable assumption that her father has any relevant authority). Or, she could observe others' behavior (Bandura and Walters 1977). Either of these strategies would presumably avoid likely costs of limit-testing, such as punishment or reputational harm. Why don't children rely exclusively on these less fraught methods? Possibly because personal experience of moral failure is an especially effective or powerful form of learning.

Precisely this argument has been made to explain adolescent limit-testing. It is argued that adolescents test the limits set around them in order to acquire a visceral understanding of the consequences of violating those limits—despite already having an abstract understanding of those consequences. For instance, they may know that insulting a friend could do lasting harm to a relationship, and yet not have the "feel" for it. Once this feel is acquired, risk-taking is reduced (Baird 2008; Rivers 2009).

This rationale for limit-testing in early childhood is less discussed in the literature. Nevertheless, this period of development is characterized by frequent conflicts between parents and children over misbehavior. As reviewed by Dahl and Killen (2018), "Naturalistic studies have found that conflicts about prohibited behaviors can occur 10 or more times *per hour* in the second year" (emphasis added). Although it is possible that this is a remarkable design failure in the moral lives of young children, a more plausible conclusion is that the failure is, in fact, part of the design.

Finally, a distinct benefit of limit-testing behavior may be the value of learning to recover from moral failures. While learning the rules that govern the world and the costs of breaking them (punishment and guilt/shame) is probably the largest portion of the benefit, there is additional value to be had from an ability to minimize costs once the violation has occurred.

6. Can We Engineer Adaptive Failure?

When moral failure happens, it can promote moral learning. But this alone does not imply that we should *create* opportunities for moral failure. There are obvious

and weighty risks associated with nudging someone towards immoral actions. Still, our analysis offers some potential strategies to mitigate those risks. We next describe several necessary conditions for achieving adaptive moral failure—ingredients that could, in theory, transform lemons into lemonade.

6.1. Highlight Failure

For the experience of moral failure to promote future moral behavior, it is necessary that the failure is personally acknowledged as such. And, indeed, people often do feel guilty without any prompting. In such cases they may be reinforcing existing moral attitudes. Yet, a variety of self-protective motivations bias people away from acknowledging their own moral failures, even to themselves (Kunda 1990). Thus, one key ingredient to produce useful moral learning from failure is to help prompt individuals to recognize that they have failed.

This may be especially for children, who are still building a basic set of internalized moral norms: Some form of external feedback could help to ensure that they encode their behavior as a moral failure, when appropriate. This could take several different forms, ranging from explicit censure to a more subtle, public reminder of the relevant rule.

6.2. Attach Failure to Acts, Not People

Guilt tends to produce more desirable learning outcomes than shame. And, among the varieties of shame, a focus on specific self-defect is associated with more desirable learning outcomes than a focus on general self-defect. Thus, moral learning from failure will work best when the source of failure is not perceived by wrongdoers as an essential property of their selves. Situations where the wrongdoer does not perceive the failure to stem from an immutable property of themselves are *recoverable*: the agent believes subsequent changes to behavior based on learning from the failure are possible, and that reasonable observers would update their expectations—would be willing to forgive—based on a change in behavior.

6.3. Aim for Moderate Failures

Failures can be small or large, and ideally should be neither. It is obvious that a failure could be so minor that it prompts no learning. At the opposite extreme, however, a major moral failure may motivate withdrawal behavior, even if the failure is only weakly indicative of the agent's character. For instance, being late frequently may only be moderately indicative of an agent's character, but being late to one's own wedding may be unrecoverable.

6.4. Pedagogy in Practice

How could moral failure be useful in pedagogical contexts? Our discussion is speculative, and certainly not prescriptive. As we have already said, there is tremendous uncertainty regarding how, when, and whether the potential benefits of moral failure could ever outweigh its costs. To be clear: Although we are discussing the possibility of creating opportunities for moral failure, we do not endorse it.

The most extreme possibility we consider is that teachers could deliberately create a situation in which they hope to prompt a pedagogically productive form of actual moral failure. This, for instance, appears to be the approach adopted in at least some elements of military basic training. An active intervention could take many different forms, such as exercises that guarantee failure in one or more pupils, or changes that merely provide the opportunity for pupils to fail morally.

Passive methods offer a milder approach. Many social environments are engineered to prevent any possibility of moral failure as far as can be achieved. Schools have zero-tolerance policies, parents maximize safety and minimize hurt to others, and companies harshly punish even small deviations from safety protocols. Thus, moral failures could be prompted not by intervening to cause moral failure, but instead by relaxing current restrictions in order to allow it to arise naturally. For instance, passive intervention in a classroom of young children might involve eliminating the prohibition of "roughhousing" styles of play. Of course, when failures in such a context do occur, it may often require active intervention by the teacher to help a child draw the appropriate lessons, as described above.

A still milder use of moral failure is to draw lessons from observational learning. This method depends on the assumption that some of the same learning processes engaged when actually failing oneself can also be engaged by observing and considering the failure of somebody else. This might occur when, for instance, observing another person who has failed triggers an act of vivid simulation or imagination of what it would be like to be that person. At a practical level, it implies a curriculum of moral education that complements abstract reasoning with concrete narratives, and narratives of moral exemplars with narratives of moral "counter-exemplars." Additionally, there is the prospect of group shame and guilt, which might lead to positive outcomes. Brown et al. (2008), Mazziotta et al. (2014), and others point out that collective guilt can lead to increased prosocial behavior, which suggests that collective guilt may also have a role to play in the active use of moral failure in pedagogy.

Along similar lines, people might learn from exposure to fictional or hypothetical moral failure—for instance, by reading literature or being prompted to imagine how they would feel if they acted wrongly themselves. Here again, the approach depends on the assumption that imagination furnishes a kind of simulated experience that engages learning mechanisms ordinarily reserved exclusively for direct experience.

6.5. Can the Ends Justify the Means?

The potential costs of encouraging moral failure are obvious and perhaps insurmountable. To begin with, according to some moral theories, it is never permissible to contribute to one moral wrong in order to prevent others (reviewed in Fischer and Ravizza 1992). Even if such cost-benefit trade-offs can be permissible, the question remains whether the benefits really *do* outweigh the costs. In this regard, three mechanisms that we have mentioned seem especially promising. The first is to ensure that, when moral failure occurs spontaneously, an effort is made to ensure that the subsequent teaching and learning processes that occur maximize the likelihood of productive moral growth: moral failures do happen, and we can design pedagogical environments to take advantage of failures rather than to merely deal with them. The second is to investigate the potential role of mere imagined moral failure (e.g. in the context of reading literature or history) to elicit some of the same adaptive mechanisms that likely accompany actual failure. Finally, there is the prospect of group shame and guilt (Brown et al. 2008; Mazziotta et al. 2014).

The value of moral failure will also depend greatly on whether lessons learned from it are enduring. There is some cause for optimism: Tangney et al. (2014) report that prisoners who have feelings of guilt about actions are less likely to reoffend than prisoners who feel less guilt, implying an effect that may last months or years. However, they also report more mixed results with shame. In particular, they report little overall effect of shame on recidivism. This may reflect the grouping of both specific self-defects and global defects into one category of "shame," though there is a significant effect of shame on recidivism when it is accompanied by externalization of blame. Of course, these correlational findings cannot by themselves rule out the possibility that the effect is driven by shame- or guilt-proneness overall in an agent, but it is consistent with mediation by the extent to which an act is indicative of character.

7. Conclusions

Errors are an important source of learning, and educators often exploit this fact. Failing helps to tune our sense of balance; Newtonian mechanics sticks better when we witness the failure of our folk physics. We consider the possibility that moral failure may also prompt especially strong or distinctive forms of learning. First, and with greatest certainty, humans are designed to learn from moral failure through the feeling of guilt. Second, and more speculatively, humans may be designed to experience moral failures by "testing limits" in a way that ultimately fosters an adaptive moral character. Third—and highly speculatively—there may be ways to harness learning by moral failure in pedagogical contexts. Minimally, this might occur by imagination, observational learning, or the exploitation of spontaneous wrongful acts as "teachable moments."

References

Agerstrom, J., F. Bjorklund, and R. Carlsson. 2012. "Emotions in Time: Moral Emotions Appear More Intense with Temporal Distance." *Social Cognition* 30(2): 181-98. <https:// doi.org/[[http://dx.doi.org/101521][dx.doi.org soco2012302181]] >

Baird, A. 2008. "Adolescent Moral Reasoning: The Integration of Emotion and Cognition." In W. Sinnott-Armstrong (ed.), *The Neuroscience of Morality: Emotion, Brain Disorders, and Development,* 323-43. Cambridge, MA: MIT Press.

Bandura, A., and R. H. Walters. 1977. *Social Learning Theory,* vol. 1. Englewood Cliffs, NJ: Prentice Hall.

Baumeister, R. F., A. M. Stillwell, and T. F. Heatherton. 1995. "Personal Narratives about Guilt: Role in Action Control and Interpersonal Relationships." *Basic and Applied Social Psychology* 17(1-2): 173-98.

Berger, J. et al. (eds) 2002. *New Directions in Contemporary Sociological Theory.* Lanham, MD: Rowman & Littlefield.

Bolle, F., Y. Breitmoser, J. Heimel, and C. Vogel. 2011. "Multiple Motives of Pro-social Behavior: Evidence From the Solidarity Game." *Theory and Decision* 72(3): 303-21. <https://doi.org/10.1007/s11238-011-9285-0>

Boureau, Y. L., and P. Dayan. 2011. "Opponency Revisited: Competition and Cooperation Between Dopamine and Serotonin." *Neuropsychopharmacology* 36(1): 74-97.

Brown, R., R. Gonzalez, H. Zagefka, J. Manzi, and S. Cehajic. 2008. "Nuestra Culpa: Collective

Guilt and Shame as Predictors of Reparation for Historical Wrongdoing." *Journal of Personality and Social Psychology* 94(1): 75-90. <https://doi.org/10.1037/0022-3514.94.1.75>

Buckholtz, J. W., M. T. Treadway, R. L. Cowan, N. D. Woodward, S. D. Benning, R. Li, and C. E. Smith. 2010. "Mesolimbic Dopamine Reward System Hypersensitivity in Individuals with Psychopathic Traits." *Nature Neuroscience* 13(4): 419-21.

Bybee, J. 1997. *Guilt and Children.* Cambridge, MA: Academic Press.

Carver, C. S. 2001. "Affect and the Functional Bases of Behavior: On the Dimensional Structure of Affective Experience." *Personality and Social Psychology Review* 5(4): 345-56.

Chiu, C., C. S. Dweck, J. Y. Tong, and J. H. Fu. 1997. "Implicit Theories and Conceptions of Morality." *Journal of Personality and Social Psychology* 73(5): 923-40. <https://doi.org/10. 1037/0022-3514.73.5.923>

Cools, R., K. Nakamura, and N. D. Daw. 2011. "Serotonin and Dopamine: Unifying Affective, Activational, and Decision Functions." *Neuropsychopharmacology* 36(1): 98-113.

Crockett, M. J. 2013. "Models of Morality." *Trends in Cognitive Sciences* 17(8): 363-6.

Crockett, M. J., L. Clark, and T. W. Robbins. 2009. "Reconciling the Role of Serotonin in Behavioral Inhibition and Aversion: Acute Tryptophan Depletion Abol-

ishes Punishment- Induced Inhibition in Humans." *Journal of Neuroscience* 29(38): 11993-9.

Crockett, M. J., J. Z. Siegel, Z. Kurth-Nelson, O. T. Ousdal, G. Story, C. Frieband, and R. J. Dolan. 2015. "Dissociable Effects of Serotonin and Dopamine on the Valuation of Harm in Moral Decision Making". *Current Biology* 25(14): 1852-9.

Cushman, F. 2013. "Action, Outcome, and Value: A Dual-System Framework for Morality." *Personality and Social Psychology Review* 17(3): 273-92.

Curtis, A. J. 2013. "Tracing the School-to-Prison Pipeline from Zero-Tolerance Policies to Juvenile Justice Dispositions Note." *Georgetown Law Journal* 102: 1251-78.

Dahl, A., & Killen, M. (2018). Moral reasoning: Theory and research in developmental science.

In J. Wixted (Ed.), *The Steven's Handbook of Experimental Psychology and Cognitive Neuroscience, Vol. 3: Developmental and Social Psychology* (S. Ghetti, Vol. Ed.)., 4th edition. New York: Wiley.

Damasio, A. R. 1994. *Descartes' Error.* New York: Random House.

Darby, R. R., A. Horn, F. Cushman, and M. D. Fox. 2017. "Lesion Network Localization of Criminal Behavior." *Proceedings of the National Academy of Sciences* 201706587.

David, B., C. Ruth, and W. David. 1993. *Using Experience For Learning.* London: McGraw-Hill Education.

Daw, N. D., and D. Shohamy. 2008. "The Cognitive Neuroscience of Motivation and Learning." *Social Cognition* 26(5): 593-620.

Dienstbier, R. A., D. Hillman, J. Lehnhoff, J. Hillman, and M. C. Valkenaar. 1975. "An Emotion-Attribution Approach to Moral Behavior: Interfacing Cognitive and Avoidance Theories of Moral Development." *Psychological Review* 82(4): 299-315. <https://doi.org/ 10.1037/h0076826>

Dienstbier, R. A., and P. O. Munter. 1971. "Cheating as a Function of the Labeling of Natural Arousal." *Journal of Personality and Social Psychology* 17(2): 208-13.

Dolan, R. J., and P. Dayan. 2013. "Goals and Habits in the Brain." *Neuron* 80(2): 312-25.

Dweck, C. S. 2008. "Can Personality be Changed? The Role of Beliefs in Personality and Change." *Current Directions in Psychological Science* 17(6): 391-4.

Finger, E. C., A. A. Marsh, D. G. Mitchell, M. E. Reid, C. Sims, S. Budhani, and D. S. Pine. 2008. "Abnormal Ventromedial Prefrontal Cortex Function in Children with Psychopathic Traits During Reversal Learning." *Archives of General Psychiatry* 65(5): 586-94.

Fischer, J. M., and M. Ravizza. 1992. *Ethics Problems and Principles.* San Diego, CA: Harcourt.

Frank, M. J., L. C. Seeberger, and R. C. O'Reilly. 2004. "By Carrot or by Stick: Cognitive Reinforcement Learning in Parkinsonism." *Science* 306(5703): 1940-43.

Gausel, N., and C. W. Leach. 2011. "Concern for Self-image and Social Image in the Management of Moral Failure: Rethinking Shame." *European Journal of Social Psychology* 41(4): 468-78. <`https://doi.org/10.1002/ejsp.803`>

Gausel, N., C. W. Leach, V. L. Vignoles, and R. Brown. (2012). "Defend or Repair? Explaining Responses to In-Group Moral Failure by Disentangling Feelings of Shame, Rejection, and Inferiority." *Journal of Personality and Social Psychology* 102(5): 941-60.

Gershman, S. J., A. B. Markman, and A. R. Otto. 2014. "Retrospective Revaluation in Sequential Decision Making: A Tale of Two Systems." *Journal of Experimental Psychology: General* 143(1): 182-94.

Gilbert, P., J. Pehl, and S. Allan. 1994. "The Phenomenology of Shame and Guilt: An Empirical Investigation." *British Journal of Medical Psychology* 67(1): 23-36.

Glimcher, P. W. 2011. "Understanding Dopamine and Reinforcement Learning: The Dopamine Reward Prediction Error Hypothesis." *Proceedings of the National Academy of Sciences* 108 (Supplement 3): 15647-54.

Haslam, N. 2004. "Essentialist Beliefs about Personality and Their Implications." *Personality and Social Psychology Bulletin* 30(12): 1661-73. <https://doi.org/10.1177/014616720

Heckathorn, D. D. 1988. "Collective Sanctions and the Creation of Prisoner's Dilemma Norms." *American Journal of Sociology* 94(3): 535-62.

Ho, M. K., M. L. Littman, F. Cushman, and J. L. Austerweil. 2015. "Teaching with Rewards and Punishments: Reinforcement or Communication?" CogSci. <https://cushmanlab.fas.har vard.edu/docs/Ho_etal_2015 .pdf>

Ho, M. K., M. Littman, J. MacGlashan, F. Cushman, and J. L. Austerweil. 2016. "Showing Versus Doing: Teaching by Demonstration." *Advances in Neural Information Processing Systems* 29: 3027-35.

Hooge, I. de. 2014. "The General Sociometer Shame: Positive Interpersonal Consequences of an Ugly Emotion." <http://repub.eur.nl/pub/51671/>

Hooge, I. E. de, R. M. A. Nelissen, S. M. Breugelmans, and M. Zeelenberg. 2011. "What Is Moral About Guilt? Acting 'Prosocially' at the Disadvantage of Others." *Journal of Personality and Social Psychology* 100(3): 462-73.

Hooge, I. E. de, M. Zeelenberg, and S. M. Breugelmans. 2007. "Moral Sentiments and Cooperation: Differential Influences of Shame and Guilt." *Cognition and Emotion* 21(5): 1025-42. <https://doi.org/1o.1o8o/o269993o6oo98o874>

Hooge, I. E. de, M. Zeelenberg, and S. M. Breugelmans. 2o1o. "Restore and Protect Motivations Following Shame." *Cognition and Emotion* 24(1): 111-27. <https://doi.org/1o.1o8o/ o269993o8o2584466>

Janoff-Bulman, R., S. Sheikh, and S. Hepp. 2oo9. "Proscriptive Versus Prescriptive Morality: Two Faces of Moral Regulation." *Journal of Personality and Social Psychology* 96(3): 521-37.

Koenigs, M., L. Young, R. Adolphs, D. Tranel, F. Cushman, M. Hauser, and A. Damasio. 2oo7. "Damage to the Prefrontal Cortex Increases Utilitarian Moral Judgements." *Nature* 446(7138): 9o8-11.

Kolb, D. A. 2o14. *Experiential Learning: Experience as the Source of Learning and Development.* Upper Saddle River, NJ: FT Press.

Kunda, Z. 199o. "The Case for Motivated Reasoning." *Psychological Bulletin* 1o8(3): 48o—98.

Leach, C. W., and A. Cidam. 2o15. "When Is Shame Linked to Constructive Approach Orientation? A Meta-analysis." *Journal of Personality and Social Psychology* 1o9(6): 983—1oo2. <https://doi.org/1o.1o37/pspaooooo37>

LeDoux, J. 2oo3. "The Emotional Brain, Fear, and the Amygdala." *Cellular and Molecular Neurobiology* 23(4-5): 727-38.

Li, J., D. Schiller, G. Schoenbaum, E. A. Phelps, and N. D. Daw. 2o11. "Differential Roles of Human Striatum and Amygdala in Associative Learning." *Nature Neuroscience* 14(1o): 125o-52.

Lickel, B., K. Kushlev, V. Savalei, S. Matta, and T. Schmader. 2o14. "Shame and the Motivation to Change the Self." *Emotion* 14(6): 1o49-61. <https://doi.org/1o.1o37/aoo38235>

Lindsay-Hartz, J., J. de Rivera, and M. F. Mascolo. 1995. "Differentiating Guilt and Shame and Their Effects on Motivation." In J. P. Tangney and K. W. Fischer (eds), *Self-Conscious Emotions: The Psychology of Shame, Guilt, Embarrassment, and Pride,* 274-3oo. New York: Guilford Press.

Lough, S., C. M. Kipps, C. Treise, P. Watson, J. R. Blair, and J. R. Hodges. 2oo6. "Social Reasoning, Emotion and Empathy in Frontotemporal Dementia." *Neuropsychologia* 44(6): 95o-58.

Lukes, S. 1965. "Moral Weakness." *Philosophical Quarterly* 15(59): 1o4-14. <https:// doi.org/ 1o.23o7/221821o>

Mazziotta, A., F. Feuchte, N. Gausel, and A. Nadler. 2o14. "Does Remembering Past Ingroup Harmdoing Promote Postwar Cross-group Contact? Insights From a Field-Experiment in Liberia." *European Journal of Social Psychology* 44(1): 43-52. <https://doi.org/1o.1oo2/ ejsp.1986>

Merrill, J., and A. E. Gross. 1969. "Some Effects of Guilt on Compliance." *Journal of Personality and Social Psychology* 11(3): 232-9. <https://doi.org/1o.1o37/hoo27o39>

Monteith, M. J. 1993. "Self-Regulation of Prejudiced Responses: Implications for Progress in Prejudice-Reduction Efforts.' *Journal of Personality and Social Psychology* 65(3): 469.

Monteith, M. J., L. Ashburn-Nardo, C. I. Voils, and A. M. Czopp. 2oo2. "Putting the Brakes on Prejudice: On the Development and Operation of Cues for Control." *Journal of Personality and Social Psychology* 83(5): 1o29.

Mosher, D. L. 1965. "Interaction of Fear and Guilt in Inhibiting Unacceptable Behavior." *Journal of Consulting Psychology* 29(2): 161.

Murphy, J. B. 2o15. "Does Habit Interference Explain Moral Failure?" *Review of Philosophy and Psychology* 6(2): 255-73. <https://doi.org/1o.1oo7/s13164-o14-o22o-5>

Nelissen, R. M. A., S. M. Breugelmans, and M. Zeelenberg. 2o13. "Reappraising the Moral Nature of Emotions in Decision Making: The Case of Shame and Guilt." *Social and Personality Psychology Compass* 7(6): 355-65. <https://doi.org/1o.1111/spc3.12o3o>

Olsson, A., K. I. Nearing, and E. A. Phelps. 2007. "Learning Fears by Observing Others: The Neural Systems of Social Fear Transmission." Social Cognitive and Affective Neuroscience 2(1): 3-11.

Olsson, A., and E. A. Phelps. 2004. "Learned Fear of 'Unseen' Faces after Pavlovian, Observational, and Instructed Fear." *Psychological Science* 15(12): 822-8.

Olsson, A., and E. A. Phelps. 2007. "Social Learning of Fear." *Nature Neuroscience* 10(9): 1095-1102.

Pagliaro, S. 2012. "On the Relevance of Morality in Social Psychology: An Introduction to a Virtual Special Issue." *European Journal of Social Psychology* 42(4): 400-05. <https://doi.org/ 10.1002/ejsp.1840>

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

Prentice, D. A., and D. T. Miller. 2007. "Psychological Essentialism of Human Categories." *Current Directions in Psychological Science* 16(4): 202-6.

Rand, D. G., A. Peysakhovich, G. T. Kraft-Todd, G. E. Newman, O. Wurzbacher, M. A. Nowak, and J. D. Greene. 2014. "Social Heuristics Shape Intuitive Cooperation." *Nature Communications* 5: 3677.

Rangel, U., and J. Keller. 2011. "Essentialism Goes Social: Belief in Social Determinism as a Component of Psychological Essentialism." *Journal of Personality and Social Psychology* 100(6): 1056-78. <https://doi.org/10.1037/a0022401>

Rees, J. H., S. Klug, and S. Bamberg. 2014. "Guilty Conscience: Motivating Pro-environmental Behavior by Inducing Negative Moral Emotions." *Climatic Change* 130(3): 439-52. <https:// doi.org/10.1007/s10584-014-1278-x>

Regan, J. W. 1971. "Guilt, Perceived Injustice, and Altruistic Behavior." *Journal of Personality and Social Psychology* 18(1): 124-32. <https://doi.org/10.1037/h0030712>

Rivers, S. E., V. F. Reyna, and B. Mills. 2008. "Risk Taking Under the Influence: A Fuzzy-Trace Theory of Emotion in Adolescence." *Developmental Review* 28(1): 107-44.

Rosthal, R. 1967. "Moral Weakness and Remorse." *Mind* 76(304): 576-9.

Ruben, B. D. 1999. "Simulations, Games, and Experience-Based Learning: The Quest for a New Paradigm for Teaching and Learning." *Simulation and Gaming* 30(4): 498-505. <https://doi. org/10.1177/104687819903000409>

Ruff, C. C., and E. Fehr. 2014. "The Neurobiology of Rewards and Values in Social Decision Making." *Nature Reviews Neuroscience* 15(8): 549.

Seymour, B., J. P. O'Doherty, M. Koltzenburg, K. Wiech, R. Frackowiak, K. Friston, and R. Dolan. 2005. "Opponent Appetitive-Aversive Neural Processes Underlie Predictive Learning of Pain Relief." *Nature Neuroscience* 8(9): 1234.

Shenhav, A., and J. D. Greene. 2010. "Moral Judgments Recruit Domain-General Valuation Mechanisms to Integrate Representations of Probability and Magnitude."*Neuron* 67(4): 667-77.

Sinnott-Armstrong, W. 2005. "You Ought to Be Ashamed of Yourself (When You Violate an Imperfect Moral Obligation)." *Philosophical Issues* 15(1): 193-208.

Sinnott-Armstrong, W. (ed.). 2008. *Moral Psychology.* Cambridge, MA: MIT Press.

Tangney, J. P. 1995. "Recent Advances in the Empirical Study of Shame and Guilt." *American Behavioral Scientist* 38(8): 1132-45.

Tangney, J. P. (2002). "Perfectionism and the Self-Conscious Emotions: Shame, Guilt, Embarrassment, and Pride." In G. L. Flett and P. L. Hewitt (eds), *Perfectionism: Theory, Research, and Treatment,* 199-215. New York: American Psychological Association.

Tangney, J. P., J. Stuewig, and A. G. Martinez. 2014. "Two Faces of Shame: The Roles of Shame and Guilt in Predicting Recidivism." *Psychological Science* 25(3): 799-805.

Tangney, J. P., J. Stuewig, and D. J. Mashek. 2007. "Moral Emotions and Moral Behavior." *Annual Review of Psychology* 58(1): 345-72. <https://doi.org/10.1146/annurev.psych.56

Tannenbaum, J. 2006. "Emotional Expressions of Moral Value." *Philosophical Studies* 132(1): 43-57. <https://doi.org/1o.1oo7/s11o98-oo6-9o56-x>

Tannenbaum, J. 2o15. "Mere Moral Failure." *Canadian Journal of Philosophy,* 45(1): 58-84.
<https://doi.org/1o.1o8o/oo455o91.2o14.997334>

Thero, D. P. 2oo6. *Understanding Moral Weakness.* Amsterdam: Rodopi.

van der Toorn, J., N. Ellemers, and B. Doosje. 2o15. "The Threat of Moral Transgression: The

Impact of Group Membership and Moral Opportunity." *European Journal of Social Psychology* 45(5): 6o9—22. <https://doi.org/1o.1oo2/ejsp.2119>

11. Risking Belief[14]

John Schwenkler

1. Introduction

It seems safe to assume that you're reading this chapter because you think it a good idea to do so. Less safe, perhaps, is the assumption that you expect to learn something from it—though I'll flatter myself in thinking that you regard this as a real possibility. There are at least different forms that this learning could take. Maybe my arguments will lead you to consider a question that you'd never before had an opinion on, and you'll come to know something about it. In that case you'll gain something valuable, and the only thing you'll pay for it is your time and attention—unless, perhaps, your mind is currently at capacity and the new knowledge pushes something else out of the way. Somewhat less likely, at least if my own past is any guide, is the possibility that I will convince you that one of your current opinions is mistaken, and you'll come to believe (or have greater confidence in) the opposite, or at least no longer believe (with the same confidence) the thing you once did. In that case the bargain is even better: you put in your time, get back some insight, and get some garbage taken out too. Either way, it seems rational to go ahead and trust that learner's instinct: you'll be better off at the end.

Or will you? Sometimes doing philosophy has the effects I just described. The things we read are rigorous and insightful, and they prompt us to challenge our preconceived opinions and come to a deeper understanding of the matters under consideration. Other times, however, the effects are less beneficial. When a philosopher's conclusions are mistaken, or her arguments invalid, then if we are moved in the direction of her position it may be by means of mere persuasion, rather than learning, that this happens, and the process may result in our understanding things less well than we did beforehand. (There are, of course, also the many times when doing philosophy has no effect on our beliefs at all.) How can you anticipate which way the present experience is going to turn out? And how should this anticipatory judgment affect your assessment of whether or not to keep on reading?

[14] John Schwenkler, Risking Belief In: Becoming Someone New: Essays on Transformative Experience, Choice, and Change. Edited by: Enoch Lambert and John Schwenkler, Oxford University Press (2020). © John Schwenkler.
DOI: 10.1093/oso/9780198823735.003.00012

Here is an answer that you, as a professed lover of wisdom, likely think you simply cannot give to those questions: that if you were to anticipate that my conclusions are false, this would give you a prima facie reason to stop reading this chapter and take up some other activity instead, so that my arguments didn't impoverish your epistemic situation. You, a philosopher, will likely regard this position as counseling a decidedly unphilosophical form of dogmatic closed-mindedness—and if there is anything that marks a philosopher, it is the way she is open to correction, engages all comers, and follows the argument where it leads (Kelly 2011). How can you do any of this, if you close yourself off to an argument just because you don't agree with where it is going? But then—wouldn't you have a reason *not* to keep reading this chapter, if that is what it is going to convince you is sometimes a reasonable thing to do? And this is, in fact, exactly what I mean to attempt.

2. The Puzzle of Doxastic Transformation

Let's take a step back. Define a *conversion* as an event that satisfies the following criteria:

- It brings about a *change in some attitude* or attitudes[1] of a person.
- The changed attitude is, or essentially involves, *commitment* to the world's being a certain way (where one's own life is part of "the world" in the relevant sense). The principal contrast I have in mind is with the non-committal states of credence, preference, and desire: thus belief and intention are paradigmatic commitments on my way of thinking, though in this chapter I'll consider only the case of ("theoretical"[2]) belief. Related puzzles about the revisability of intention have been treated extensively elsewhere in the literature.[3]
- The changed attitude is *central to the person's worldview or way of life.*
- There is a considerable *distance between* the new attitude and the old one.

One of the things that makes conversion philosophically interesting is the difficulty in seeing how an event that satisfies the above criteria could consist in anything more than a *brute,* or extra-rational, change in the attitude or attitudes in question.[4] As Bas van Fraassen has observed, the questions that arise here are similar to classic questions

[1] For the sake of simplicity, I'll mostly speak in what follows of a singular "attitude" that is the locus of conversion.

[2] See Marusic and Schwenkler (2018) for the distinction between practical and theoretical belief.

[3] For a start, see the essays collected in Bratman (2018).

[4] This is roughly the problem that John McDowell raises in his important paper "Can there be external reasons?," which has had a significant impact on my thinking about these matters. McDowell introduces the term *conversion* as a placeholder for "the idea of an intelligible shift in motivational

about the rationality of scientific revolution: if attitude (or theory) Y appears simply *ludicrous* by the lights of attitude (or theory) X,[5] and X explains well enough most of the phenomena that Y is supposed to account for, then it can be difficult to see what *rational* means there could be of getting a person committed to X to abandon this commitment and commit herself to Y instead. While I think these questions are very important, my focus in this chapter will be somewhat different. I will take it for granted that it *is* possible for conversion to take place through the proper operation of one's rational faculties rather than by means of brute force or some other non-rational process. And the question that I am going to raise concerns whether, and in what circumstances, the seeming falsity of a view could reasonably be counted as a reason not to take the risk of engaging with arguments in its favor, or with experiences with the potential to convert one to it, since by one's current lights the *result* of having one's mind changed by those arguments or experiences would be, as L. A. Paul puts it, a kind of *cognitive impairment*:

> When is it the case that, before I make a decision to become a different sort of self, I can rightly regard my future self as cognitively impaired, relative to my current self? Should I, from the perspective of my current self with her current preferences, regard any dramatic change of my preferences, especially transformative changes to my core personal preferences, as a kind of cognitive impairment? Where is the line between revising one's preferences in response to experience such that one autonomously learns from the experience, versus having one's preferences controlled by the experience? (Paul 2015: 806)

While Paul frames her puzzle in terms of a change in one's personal preferences, a similar worry arises when the attitudes that stand to be changed by an experience are *commitments* in the sense defined above. Suppose, for example, that you believe that God exists, that government should represent the will of the people, or that it is wrong to eat the flesh of animals when meatless options are readily available. And now suppose that I propose to present an argument, or expose you to some sort of experience, that has a good chance of getting you to abandon this belief, and perhaps to believe the opposite thing instead. By your current lights, the transformation I'd thereby bring about would be a cognitive impairment, not an improvement. And while of course it is true that, by the lights of the person you'd be in the wake of this transformation, the change would have improved your cognitive situation quite a lot, by your current lights it seems to you that this is just what you *would* think, as the

---

orientation that is exactly *not* effected by inducing a person to discover, by practical reasoning controlled by existing motivations, some internal reasons that he did not previously realize he had" (McDowell 2001: 102).

[5] If, that is, "the new view is literally absurd, incoherent, inconsistent, obviously false, or worse—meaningless, unintelligible—within the older view" (Van Fraassen 2004: 72).

meat-eating, democracy-hating atheist you would unfortunately have become.[6] In such a situation it seems reasonable to ask: is that really the sort of thing that you should risk having happen to you? Doesn't the perceived badness of the outcome give you at least *some* reason to keep it from coming about?

The *puzzle of doxastic transformation* that I'm trying to get you to take seriously has a similar structure to the puzzle of personal transformation that Paul puts forward in *Transformative Experience* (Paul 2014). Paul's puzzle is supposed to arise when (1) a person has to make a choice that is based on considerations of her own subjective well-being, but (2) since the choice involves a sort of experience that she's never had before, she's not in a position to weigh its value or determine how she is likely to respond to it. Moreover, (3) the person knows that if the experience does

alter her preferences, then she'll be satisfied living in a way that would not satisfy her at all presently. And what makes Paul's puzzle puzzling is the intuition that, given the limitations in (2) and the asymmetry in subjective satisfaction that arises due to (3), the person isn't able to choose wisely as per (1).

As I noted just above, an important difference between my puzzle and Paul's is that in the cases that interest me a person's core beliefs are at stake, and not merely her core preferences (though of course a conversion might alter those too). But by my lights this makes the puzzle of doxastic transformation only more pressing, since beliefs aren't subjective in the way that (at least some) preferences are. Because of this, the worry that through some transformative event I may end up preferring things wildly different than the ones I currently prefer is not as serious as the worry that I will end up getting things wrong by coming to have a lot of false beliefs. My puzzle, then, arises when (1') a person has to make a choice that is based on considerations of her own cognitive well-being, but (2') she's not in a position to say in advance whether the choice will lead to an improvement in her cognitive situation rather than an impairment of it. Moreover, (3') the person knows that if her beliefs were to change, whether rationally or not, then she'd regard this change as an improvement. And so the puzzle concerns the fact that, given the limitations in (2') and the asymmetry in cognitive perspective that arises due to (3'), the perspective from which the person must make the choice in (1') appears inadequate to the task. At least as far as epistemic goods are concerned, the rational choice to make is the one most likely to lead to true beliefs (or knowledge

[6] Again, van Fraassen puts it well: "how can we tell that what we see as material welfare in that kind of future will be cognized as such then? And if it is not, what about a future in which we are by our present lights well off, and by our lights then miserable or suffering a great loss? Perhaps we can suppose that in contemplating the decision I see myself laughing and outwardly cheerful in that future. But given the opacity to me now of what my words and body language then really mean, I really have no access to how things really seem to me then. Will this cheerfulness be the false face adopted in despair of ever regaining what I have lost? None of these difficulties appear in the ordinary everyday case where we can assume that our future way of thinking about ourselves will be the same as it is now, factual details aside. That, however, is precisely what we cannot assume here. There is therefore no rational way of deciding upon such a transition, if rational means 'rational by the lights we have beforehand' " (2004:101-2). This last conclusion (following "therefore") is one I want to resist.

or understanding or wisdom, etc.—I take no standing on the hierarchy of epistemic value) by way of a rational process. But the person making the choice is barred from scrutinizing directly the epistemic standing of the potentially transformative process, and her attempt to evaluate which beliefs are true and which belief-forming processes are epistemically upstanding will necessarily reflect her current perspective.

Perhaps you think that this puzzle is not really puzzling at all, or that you have on hand an easy solution to it. In that case I invite you to go ahead, take the risk. Let's see what I can do to change your mind.

3. An Objection: The Whole Thing Is Badly Conceived

"Choosing what to believe? This reeks of doxastic voluntarism. And what's with this business about treating the potential truth and falsity of our beliefs as relevant considerations in our decision-making? Has anyone actually done this, ever? All this feels like a philosopher's pseudo-problem: irrelevant to everyday life, with its supposed force dependent on a questionable framing and a bunch of unsupported assumptions."

Hold on there. I agree that doxastic voluntarism is a Bad Thing—and if anything in my setup of the puzzle of doxastic transformation presupposed it, then this would indeed be sufficient reason to think the puzzle ill-conceived. But I haven't presupposed any such thing. The doxastic voluntarist holds that belief itself is subject to the will—that is, that it's possible to choose, or at least to exercise some degree of voluntary control over, our doxastic commitments themselves, in something like the way that we exercise voluntary control over our bodily movements. And what makes this impossible is that it's in the nature of belief to "aim at truth," and thus to be responsive only to considerations that bear on the truth or falsity of the matter in question, whereas whether one *wills* or *wishes* to believe something isn't usually such a consideration at all. This doesn't, however, mean that our doxastic situation isn't to a great extent the product of voluntary control, since the choices we make about such things as what to watch or read, what to study and where to do it, whom to be in conversation with and about what topics—all of which concern matters that are clearly under the authority of the will—have an undeniable influence on what we go on to know and believe.[7] Care about who won the Knicks game last night? Turn on SportsCenter. Interested in international affairs instead? Pick up the *New York Times*. Want to learn more about the forces that govern the interaction of subatomic particles? Maybe you should take a course in physics. None of this implies that you can learn about these things, or influence in any way what you believe about them, simply by some internal act of the will.

[7] All these are examples of what Pamela Hieronymi (2006) helpfully calls "managerial control" of one's attitudes.

The second thing is trickier, since I *do* want the puzzle I am raising to be anchored in questions that concern us in the course of everyday life, and it is fair to complain that the problem I've outlined does not capture very well the "lived" structure of any real-life quandaries. That is, however, partly because this representation of the puzzle *abstracts* from so much of what we take to be relevant in making choices with the potential to influence our future beliefs—considerations like wealth, power, comfort, convenience, the desires of our loved ones, and so on. And this abstraction has a purpose, namely to shift our attention away from those considerations and toward the question that is my focus here, i.e. that of how (if at all) we should regard the value of true belief *itself* as a factor in our practical deliberations. I think it is clear that most of us *do* care, to some degree at least, simply about getting things right, and we are disposed to make choices that help us to do this. Doesn't this generate rational pressure to avoid making choices that threaten to leave us epistemically worse off?

This is not to say that the risk of getting things wrong usually presents itself as such as a salient factor in practical deliberation. But *sometimes* it does. Think, for example, of the secret agent who is tasked with infiltrating an extremist group whose ideology she thoroughly rejects, and who must account for the possibility that the time she spends in close contact with these extremists will soften her opposition to their beliefs. A similar situation may arise for an academic researcher who is interested in studying the effects of propaganda or media bias: isn't it sensible for her to weigh the risk that in exposing herself to a barrage of the messages whose rhetorical force she wishes to investigate, *she'll* be influenced by those messages to some degree, and emerge from her research having "learned" that they contain a good deal of truth? And then there's the case of the dogmatic religious fundamentalist, who worries that her (or her child's) salvific worldview could be harmed by exposure to a too- sympathetic presentation of the supposed discoveries of modern science. *That* is an instance of closed-minded anti-rationalism, you are likely inclined to say. And I agree. But is this simply because of the way it involves regarding dramatic change in one's core beliefs as a kind of cognitive impairment?

4. Four Inadequate Responses to the Puzzle

4.1. Stand Pat

Let's start by considering the position of our imaginary dogmatist (who is, it should be emphasized, merely a philosopher's caricature of her real-life counterparts). In deciding whether or not to make a choice that promises to change a certain core belief of hers, the dogmatist treats the truth of that belief as a fixed point in her practical reasoning. As long our dogmatist reasons in this way, the prospect of a changed belief is guaranteed to be, not ruled out (for there may be further considerations that

speak sufficiently in favor of openness to the possible transformation), but regarded as something that should *so far* be avoided.

As I've just indicated, it doesn't follow from this way of thinking that the dogmatist will *never* choose in favor of potential doxastic transformation. Sometimes she might make such a choice on the basis of non-epistemic considerations, such as the value of wealth, popularity, or power, or of sharing a valuable experience with her friends or family. But it is possible as well for epistemic considerations to be taken up in this sort of calculus: while having a mistaken belief concerning a certain matter might be counted as very bad, this particular badness could be outweighed by the goodness of gaining correct beliefs about a whole lot of other things. The rational space for these choices will shrink, however, as the belief at stake becomes more central to the dogmatist's identity or way of life, since then the epistemic disvalue of losing this belief will appear much greater, and the conception of the world that is informed by this belief will assign less value to goods that are in tension with it. In the limit, the prospect of losing this belief will appear as a kind of personal extinction, such that *nothing* in the world could be worth the risk of giving it up.

Is it ever rational to take the attitude of the dogmatist toward the possibility of doxastic transformation? As I will discuss just below, I am inclined to think that sometimes it could be. It is not plausible, however, that this attitude could be the rational one to take *in general*, as it counsels a form of closed-mindedness that makes it difficult to respond appropriately to opportunities for learning. If my conclusion in this chapter is correct, it is indeed reasonable to be vigilant about the possibility of future cognitive impairment. But it seems quite clear that the situation in which one's mind is changed, even on an important matter that is central to one's identity or way of life, must not *always* be viewed in prospect as something that should so far be avoided.[8]

4.2. Risk It

The next response we'll consider says that the only rational response to the possibility of doxastic transformation is an unhesitating open-mindedness. According to this position, it is simply a mistake to approach a doxastically transformative choice by worrying about the possibility that it will leave one with a false belief—for *that* is something that one will worry about only if she takes for granted the very belief that the transformation might lead her to question or reject, and to reason in this way would beg the question at issue. By contrast, the mark of epistemic virtue is to open

[8] Though she appeals to the value of having a stable perspective over time rather than the value of true belief itself, I take Sarah Paul's (2015a; 2015b) defense of the rationality of doxastic self-control to be vulnerable to this objection. While *sometimes* the importance of diachronic continuity might provide sufficient reason to refuse to reconsider a belief in the face of countervailing evidence (or expose it to such evidence when one takes it to lie around the corner), this cannot *always* be the case, as then our dogmatist would be rational in remaining steadfastly closed-minded.

oneself up to challenging arguments and unexpected experiences, confident that the *true* position is going to emerge through a process of radically open-minded inquiry.

As attractive as this position can appear at first glance, on further examination it appears to rest on an unrealistic view of the ability of human learners to respond rationally to potential sources of new information. Following Sarah Paul (2015b), let's use the term *epistemic temptation* to describe the sort of situation in which false or unwarranted beliefs strike us as apparently reasonable. These situations are common enough that we have a special class of verbs to indicate the possibility that we may be in one: the stick in the glass of water *looks* like it is bent, the argument *appears* to be valid, she *sounds* like she is telling the truth, it *seems* like this is the correct conclusion to draw from these data. And while of course we are capable of doing a pretty good job of distinguishing reality from appearance in these domains, that ability is far from foolproof, and it is important to be aware of circumstances or domains of inquiry in which epistemic temptation is especially likely to lead us astray. To the extent that a potentially doxastic transformative event appears to have these characteristics, this counts as a prima facie reason to avoid it.

That's not, though, to recommend the line of thought that we attributed to our radical dogmatist: that since P is true, and in doing F I might be led not to believe it, therefore doing F is to this extent not recommended. For we can reject the unequivocal advice always to Risk It without simply taking it for granted that our present beliefs are true, and that *anything* that would lead us to change them is therefore a mere temptation. What stands behind our verdict is the much more modest observation that the chance of giving in to epistemic temptation is something we should take seriously in making choices that have the potential to change what we believe. The challenge is to identify how this can reasonably be done.

4.3. Proceed With Caution

The extreme responses we have considered so far—according to which we should always Stand Pat, viewing doxastic transformation always as a cognitive impairment, or always Risk It, viewing doxastic transformation always as an opportunity to improve our epistemic situation—are both inadequate. *Sometimes* a doxastically transformative event amounts to a way of learning about the world. But there are other such events that amount to regrettable cognitive impairment instead. How can we approach occasions of potential doxastic transformation in a way that maximizes the first of these outcomes over the second?

An obvious answer is that the choice for a potential doxastic transformation doesn't have to be all or nothing. In choosing to expose yourself to something—an argument, a powerful experience, a series of relationships with people whose ideology you think is false—that has the potential to change what you believe, you needn't thereby choose to allow your beliefs to be changed in this way, no matter what. For it's possible for you to enter into the potentially transformative process with your guard up, thinking

carefully about what you encounter and subjecting to rational scrutiny any inclinations you have to change what you think. Doesn't this give you a way to ensure that any doxastic transformation that you do end up undergoing constitute an improvement, rather than an impairment, in your epistemic situation?

Attractive as it may appear on a first pass, this response assumes a too-optimistic picture of the ability of human reasoners to distinguish the true from the false. First, however good you are at thinking carefully and keeping up your rational guard, there are bound to be some situations in which your thought becomes careless and your guard slips—situations in which, despite your best efforts, you nevertheless succumb to epistemic temptation. And second, to the extent that you do keep up your guard, this will often involve evaluating potentially transformative influences against the background of the very beliefs that they threaten to change. Our imaginary dogmatist, for example, can't very well count as having opened her mind to scientific discoveries if every time she encounters a scientific claim, she reasons that since it conflicts with her religious worldview, therefore it can't be true. Potentially transformative events won't teach a person anything if she approaches them so cautiously that they never have a chance of making a difference to the way she understands the world. The advice to Proceed With Caution is probably on the right track, but on its own it offers no real solution to our puzzle.

4.4. It Depends

Our first three responses all take the form of universal policies: they say that one should *always* Stand Pat, Risk It, or Proceed With Caution in the face of a potential doxastic transformation. And in considering these responses, we saw that none of these policies is *always* the right one to adopt. There's not, however, any good reason to require that choices of the sort we are considering will admit of a one-size-fits-all solution. Indeed, it seems on the contrary that the way to approach a doxastically transformative choice will *depend* on the epistemic credentials of the belief that such a transformation would lead one to reject, as well as on the nature of the process that would lead one to do this. This is part of why we think, for example, that a scholar is justified in acting to preserve her beliefs in a way that a conspiracy theorist is not: for the scholar has *good reason* to believe as she does, whereas the conspiracy theorist is a paradigm of irrationality. It is also why we think it more justifiable to close oneself off to forms of doxastic transformation that involve, as Paul puts it in the passage that I quoted in Section 2, having one's beliefs "controlled" or *manipulated* than it is to avoid opportunities for *learning* from novel experiences (which may include such things as study, argumentation, and so on). Can appeal to considerations like these show the way out of the difficulties we have raised?

Let's consider first the possibility that the crucial difference lies in the epistemic status of the belief that a doxastically transformative experience promises to change. The idea here may be that to the extent that a person is *justified* in believing something,

she will also be justified in believing that coming to believe the opposite thing will constitute a cognitive impairment. And this is because a person's justification for believing the thing in question will also be her justification for rejecting the possibility that a change to this belief would put her more in touch with how things really are. The difficulty, however, is that there is a kind of justification that can be possessed even by our imaginary dogmatist or conspiracy theorist, who may be able to appeal to any number of considerations that she takes to support her point of view. And by the same token, many of the beliefs that we seem to be the most justified in refusing to risk the loss of are also beliefs that we seem to have very little justification for: the belief in democracy, for example, or in the dignity of humanity, or in the importance of tolerance and open-mindedness. Cases like that of the justified conspiracy theorist suggests that justification for a belief may not be sufficient grounds for refusing to expose it to counterevidence. Cases of the latter sort suggest that this may not be necessary to make such a refusal rational, either. In each case, the appeal to justification is insufficient to solve our puzzle.

The other possibility we need to consider is that transformative processes can be evaluated according to the means by which they bring about doxastic transformation. For example, part of what makes it rational for you to keep on reading this chapter even if you would rather not be brought around to its conclusion—or for our imaginary dogmatist to open her mind to scientific evidence despite the worry about where it will lead her—is that encountering scientific evidence and engaging with philosophical argumentation are ways of learning how things are. This makes them different from exposure to pure propaganda or various forms of emotional manipulation, which influence or control our beliefs in ways that we regard as epistemically problematic even if the beliefs they bring us to have happen to be true. Given this distinction, can't we evaluate the epistemic credentials of potential doxastic transformations, not by considering just the truth or falsity of the beliefs that they might result in, but also by reference to the means by which they might bring these results about?

I'll argue just below that a version of this response is defensible, but for now we should notice an obvious difficulty with implementing it. The advice we are considering says that a person should evaluate an instance of potential doxastic transformation by considering whether or not it involves a process of learning rather than one of mere control or manipulation—but by what criteria, and according to which standards, is one supposed to decide this? Consider the situations of an atheist trying to decide whether to participate in an emotionally charged religious revival, and a Christian fundamentalist trying to decide whether to keep an open mind as she takes a course in evolutionary biology. In each case the deliberator's current perspective might lead her to regard the process by which her beliefs would be changed as largely non-rational or insufficiently driven by evidence: the atheist, because she thinks of religious experience as emotionally charged hallucination; and the fundamentalist, because she thinks of science as ideology and of human reason as too deeply flawed to penetrate the mysteries of creation. Yet things would appear quite differently from the perspective that each

of them would have if the transformative processes in question were to unfold. Having been brought to religious belief by a powerful experience, our former atheist will see such experience as the revelation of a transcendent reality. Having been convinced of the credentials of science by an open- minded course of study, our former fundamentalist will see the methods of scientific inquiry as an appropriate way to supplement and even reshape our religious convictions. How should these competing perspectives be prioritized? We are faced with another version of the original puzzle.

The difficulty here is not that there is no way, from the perspective of a given system of belief, to scrutinize the epistemic credentials of a process with the potential to transform that system. It is rather that in order to do this effectively one cannot simply take for granted the beliefs that this process threatens to transform. At least some of these beliefs will need to be "bracketed," as we sometimes say, in order to evaluate the credentials of the process from an appropriately neutral perspective. However, to the extent that the beliefs in question really are central to one's way of understanding the world, if they are taken out of play there may be not enough left to bring one's reason to bear on the crucial question. For there is, as van Fraassen reminds us, no such thing as an epistemology that is altogether independent of our presuppositions about the nature of the world and our situation in it:

> There is no way to write a theory of cognition while escaping from our general beliefs about what we and our world are like. We cannot construct a presuppositionless theory, a priori, independent of our current science, theology, metaphysics, or whatever else we have accepted as knowledge. But neither can we construct a theory based in our current knowledge base and still make sense of the idea that we might be ... capable of correctly attaining, through a conceptual revolution, a true insight radically at odds with that current knowledge base.
>
> (2004: 81-2)

A bit later on I will revisit this last position, attempting to identify a way in which the epistemic credentials of a doxastically transformative process can be evaluated rationally and in a non-question-begging manner. First, however, I need to dislodge an assumption that made our original problem seem so intractable.

5. A Better Way Forward

So far I've represented the puzzle of doxastic transformation in terms of a situation in which a person's current beliefs lead her to the conclusion that certain dramatic changes in those beliefs would constitute a kind of cognitive impairment. The puzzle arises from the fact that it seems at once unwise to disregard this possibility altogether and unreasonable simply to take one's present beliefs for granted in weighing the likely

costs and benefits of a potential transformation, or evaluating the epistemic credentials of the processes it would involve. Nor, however, can one approach these questions in a manner that abstains from any potentially controversial claims about what reality is like and what ways we have of coming to know it. But which of such claims is it reasonable to rely on?

The arguments in Section 4 were all focused on showing that not all the things a person believes to be true are appropriately taken for granted in reasoning about potential doxastic transformation. What I will propose now is that with knowledge, the situation is different.

A simple case will illustrate the basic idea. Imagine that you believe that you left the stove on before leaving the house, and so you return home to turn it off. And now imagine further that the stove wasn't left on after all—you turned it off when you finished cooking your meal, but have altogether forgotten having done this. The stove is not on, but your decision to go home and turn it off still makes a kind of sense: you are heading back home because you believe that the stove is on. You are not, however, going home because the stove actually *is* on—this can't be what explains your going home, since it isn't the case at all.

What more would be required for the stove's being on—the *fact that* it is on, as we say—to be what explains why you go home to turn it off? One thing, obviously, is that the stove would need to *be* on: for an explanation of the form "X because *Y*" can't get off the ground if *Y* isn't even the case. Yet the stove's being on also can't explain why you are going home unless that fact is related to your action in the right kind of way. Thus if, for example, you turned the stove off before you left the house and then it was turned back on by someone else, then the stove's being on still won't be what explains why you are going home. Here you are in the lucky position of acting from a belief that happens to correspond to how things are, but still it is merely your belief, and not the fact that things are that way, that explains what you are doing.

This sketch of an argument needs a good deal of filling in, but this isn't the place to provide that.[9] Instead I am going to take this groundwork for granted and develop what I think is the correct conclusion to draw from it, which is that *knowledge* of a fact is what makes it possible for the fact itself to be what explains, and therefore rationalizes, the things we do on the basis of it in a way that mere belief in a fact does not. And this means that someone who *knows* that something is true can respond to the value of believing this, and the disvalue of believing the opposite, in a way that someone who merely *believes* something to be true cannot.[10] In this respect we may

[9] I take the necessary groundwork to have been laid by Hawthorne and Stanley (2008) and Hyman (2015: chs 6-8), among others.

[10] Notice that in each case—for the knower as well as the mere believer—it is only the value of true belief itself that I assume to be at stake. The case for a knowledge-centered solution will get only stronger if the value of knowledge exceeds that of mere true belief—though the ability to respond rationally to that added value might depend on knowledge that one knows, which again is not assumed in my argument here.

contrast the situation of the academic researcher worried about the effects that a close study of extremist subcultures might have on her understanding of the world, with that of a paranoid skeptic who shuts out the "mainstream media" in order to keep himself from being taken in by globalist propaganda. What makes their situations different is not just that one of them happens to be correct while the other is not—for even a conspiracy theorist will sometimes stumble upon the truth. Rather, I suggest that the crucial difference is that, insofar as the researcher has *knowledge* of the worldview that she wishes to keep her research from undermining, the *facts* that she reasons from can serve as her justification for deciding against a course of action that would put her out of touch with those facts. Our skeptic, by contrast, decides as he does only because he *believes* that the world is a certain way. And the mere fact that one believes something does not provide the same kind of justification for acting in ways that maintain this belief that a known fact can provide for acting in ways that keep one cognitively in touch with it.

Let me guess at what you are thinking. Even if this position is correct as far as it goes, as a proposed solution to the puzzle of doxastic transformation it leaves a lot to be desired. That's because a mere believer, no less than a knower, will very often *take* herself to be justified in reasoning, not merely from the fact of her beliefs, but from the believed facts themselves, to a conclusion that she takes those facts to justify. This sort of thing happens *all the time*: e.g. thinking that my keys are in the kitchen drawer, I open the drawer to get them, and it's only after my search comes up empty that I'm inclined to say that I opened the drawer only because I *thought* the keys were there, and not because they actually were. So, at least in the absence of a reliable way of distinguishing knowledge from mere belief, the solution fails to provide the sort of guidance we were seeking.[11]

There are several things to say here. First of all, notice that even the advice to act *according to one's beliefs,* or those of one's beliefs that are the most justified or confidently held, would have to face its own version of this objection unless we could rule out the possibility of being mistaken as to what one's own beliefs (or one's most justified or deeply held beliefs) actually are.[12] If matters like these are not infallibly self-known, then even a philosopher who claims that it is always ("subjectively") rational to act according to one's current lights will have to allow for the possibility that a person can be wrong about what it is rational for her to do. Second, notice also that nowhere I have claimed that knowledge *of knowledge* is required for one's choices to be rationalized by a worldly fact. What I have proposed is that a person can rationally choose to (do F because P) only if she knows that P is the case, and that if P is something she merely believes and does not know, then the most she can rationally choose is to (do F because she believes that P). On this account, just as it is possible

[11] Thanks especially to Kyla Ebels-Duggan for pressing me on this point.

[12] Foran argument that *none* of our mental states are "luminous" in this way, see Williamson (2000: ch. 4).

to think that a choice would be rationalized by the facts when really the choice would be grounded merely in one's beliefs, so *if* it is possible to know something without knowing that one knows this, then it is possible to lack knowledge of what it is rational for one to do. But this doesn't mean that one can't *rationally choose to do one thing or another*—not, anyway, unless we assume that matters of (even "subjective") rationality are infallibly self-knowable, which (as we've just seen) is arguably unachievable on *any* reasonable view of the limits of human self-knowledge.[13]

This leads to a third and final point, which is that an adequate response to the puzzle of doxastic transformation does *not* require providing guidance that shows how to decide, in any given case, whether or not to act in a way that invites the possible transformation of one's core beliefs.[14] As I introduced the puzzle in Section 2, I gave it the form of a "How possible?" question: the puzzle seeks an explanation of how it *could be* rational on some occasions to resist potential conversion on the grounds that it would constitute cognitive impairment, and on others to open oneself up to possible conversion on the grounds that it provides an opportunity for learning. If this is the question we are trying to answer, then our question is answered by recognizing the special way that knowledge puts us in a position to act in light of the facts, as then we can understand the goodness of an inference like

> (K) P, so I should avoid doing F, as this would bring it about that I believe that not-P

in a way that does not carry over to an inference like

> (B) I believe that P, so I should avoid doing F, as this would bring it about that I believe that not-P.

The recognition that sometimes a person can reason along the lines of (K), and so choose rationally against potential conversion in a way that a person who reasons merely along the lines of (B) could not, speaks directly to our "How possible?" question, even if it does not answer the further question of how to tell whether one is in a position to reason in the first way rather than the second. There is of course a corresponding "How possible?" question, as well as any number of more practical ones, concerning how knowledge and mere belief can be distinguished from the first- person perspective—but these matters fall outside the scope of the present inquiry.

Let's return now to the responses to the puzzle of doxastic transformation that were discussed in Section 4.4. In that section I argued, first, that mere *justification* for a belief is neither necessary nor sufficient ground for deciding that an experience

[13] For further discussion of this last matter, see Hawthorne and Srinivasan (2013).

[14] As Nomy Arpaly put it to me, we should not demand that philosophy provide us with a *manual* that will tell us what to do, or believe, given our beliefs, preferences, and facts about the world as we find it.

that would change this belief would amount to a form of cognitive impairment. We are now in a position to see why this is. Justification for a belief, at least of the sort possessed by our hypothetical fundamentalists and conspiracy theorists, is not *sufficient* to rationally reject a doxastically transformative experience, simply because justification of this sort is fallible: it is possible to have this sort of justification without thereby *knowing* the thing that one justifiedly believes. And it will not be *necessary* for such a decision as long as justification is not necessary for knowledge— that is, as long as it is possible to know certain things, e.g. that human beings are ends in themselves, without this knowledge depending on further justification.

The other thing I argued in Section 4.4 was that a person cannot reasonably decide whether a given process of doxastic transformation would center on *learning*, rather than mere influence or control of her beliefs, in a way that simply takes for granted the beliefs that the process promises to change—as an atheist might assume that religious experience is entirely hallucinatory, or a conspiracy theorist might think that scientific consensus is a state-sponsored hoax. In addition, I argued, if a person were to "bracket" or set aside *every* part of her worldview that a process promises to change, then there might not be enough of that worldview remaining for her evaluation of the process to appeal to. The upshot was that we could not solve the puzzle of doxastic transformation simply by saying that one should evaluate the kind of process that a potential doxastic transformation would consist in.

But there was something troubling in this conclusion. For the sort of reasoning whose possibility this argument calls into question seems like a sort of reasoning that people *must* sometimes engage in, and it would be startling to discover that there is no way of engaging in it rationally. Suppose, for example, I have no idea how old the Earth is, and am considering two possible courses of study that would lead me to different conclusions about its age. If one of these courses of study centers on careful reasoning and the consideration of scientific evidence, while the other takes its bearing from influential myths and centers on patterns of inference that most experts reject, then this does seem like a good reason to favor the former course of study over the latter—even as we concede that if the conclusion of the second course of study is correct, then mythology will have been underrated, and scientific reasoning overrated, as a means of obtaining geological knowledge. Does it follow from this concession that we can't appeal to the importance of evidence-based reasoning in choosing to avoid doing things that would lead us to question this importance?

We are now in a position to respond to this challenge. What makes it possible to appeal to the importance of evidence-based reasoning in evaluating the epistemic credentials of a potentially transformative process is our knowledge that critical reasoning, and careful consideration of publicly observable evidence, are good ways of learning about things like the age of the Earth, while endorsing a literal reading of ancient myths is not. Since we know these things about what is and is not a good way to learn about the world, we can reason from these very facts to knowledge about the kinds of doxastic transformation that are likely to teach us things (i.e. to give us new

knowledge), and not merely manipulate us into believing things that might or might not correspond to the facts. In reasoning about what to do in the face of a potentially transformative process, including about the credentials of that process itself, the things one knows do not have to be bracketed.

This last point also allows us to address one more lacuna in the proposal I have put forward in this section. My concern in this chapter has been mainly with understanding the potential rationality of choices against potential doxastic transformation—that is, choices not to go in for a doxastically transformative experience, on the grounds that such an experience would worsen one's epistemic situation. I have argued that such a choice can be rational only if it is a response to the truth of the belief in question, i.e. to a fact that one knows to be the case and not merely a belief that one holds. But I haven't said anything about what might make it possible to see a doxastically transformative choice as having positive value—that is, as positive with respect to the value of the belief it would result in, and not merely the other things that might be taken to recommend it. And it is still hard to see what could account for this: after all, a doxastic transformation is supposed to bring it about that one has a belief that one currently thinks is false, so how is it possible to see such a transformation as potentially good?

I answer that one should welcome a doxastic transformation, seeing it as an occasion for improvement in one's epistemic situation, to the extent that it centers on processes that are ways of learning about how things are. If, for example, you believe but don't know that atheism is true, then on this account you do have good reason to engage with philosophical arguments with a good chance of convincing you of the existence of God, since philosophical argumentation is a way of learning how things are. But the same verdict will not hold if you are considering instead whether to start spending time with the members of an emotionally manipulative cult—not because the views of that cult are wrong (we are assuming you do not know this), but because their ways of bringing you around to those views would ways of controlling your beliefs, and not a means by which you would learn how things are. You might, of course, still have some reason to spend significant time with the cult members (perhaps they are members of your family, or you are conducting sociological research on this cult's beliefs, etc.), though likely you would do this only in the cautious manner discussed in Section 4.3, and you would *not* be doing it because of the epistemic value of the manipulation that you would be subjected to.

6. Conclusion

Should you have kept on reading this chapter, then? I hope to have shown that it depends on what you knew at the beginning. *If* you had known that the puzzle of doxastic transformation was not a puzzle at all, or that one of the responses to the puzzle that I criticized in Section 4 was an adequate response, then by my own lights

the possibility that I would convince you to believe otherwise was something that you could reasonably count in favor of putting the chapter aside. But I don't believe you *did* know that—and, indeed, I think the arguments I have made are sufficient to show that I *know* you didn't know any such thing. If I'm wrong in that, and my arguments have managed to bring you around to my position anyway, then I suppose I owe you an apology—though I do hope the experience was still enjoyable enough, and illuminating with respect to some of the subsidiary claims that I have argued for, to make up for that bit of damage done.

7. Coda

I need to add a confession of disappointment. When I first began working on this project, I was convinced that it *had* to support the conclusion that a person could never be so blinkered by a mistaken worldview that there was simply "no way out" for her—no way, that is, that she *could possibly* choose rationally in favor of an experience that threatened to convert her to something better. My reasons for this conviction were broadly religious: for the possibility of there being "no way out" seems to mean that a person could get to be so badly off that her conversion could happen only against her will—through an event which, *if* she had anticipated what it involved, she *could not* have rationally allowed to happen.

I now think that there is indeed a possibility of this sort. What grounds this possibility is *not* that it is always irrational to make choices that threaten to change our core beliefs, but rather that what it is rational for us to do depends on what we know—and a person with a radically false worldview might be too *ignorant* to reason successfully about which choices will improve her epistemic situation. To the extent that this is true of a person, the only way it can be rational for her to make a choice that would convert her is if there is something else—the invitation of a loved one, perhaps, or the fresh air of a garden, or the promise of getting to Emmaus or Damascus—that grounds her sense of its value, leading her to risk what she has and thereby end up with something much better, something whose worth she could not have recognized in advance.[15]

References

Bratman, Michael. 2018. *Planning, Time, and Self-Governance: Essays in Practical Rationality.* Cambridge, MA: Harvard University Press.

[15] For discussion of these issues, and feedback on earlier versions of this chapter, I am grateful to Nomy Arpaly, Joshua Blanchard, Lara Buchak, Nick Byrd, Nilanjan Das, Josh DiPaolo, Ravit Dotan, Trent Dougherty, Kyla Ebels-Duggan, Anne Jeffrey, John Kvanvig, Enoch Lambert, Sam Lebens, Clay-

Hawthorne, John, and Amia Srinivasan. 2013. "Disagreement Without Transparency." In David Christensen and Jennifer Lackey (eds), *The Epistemology of Disagreement: New Essays*. Oxford: Oxford University Press.

Hawthorne, John, and Jason Stanley. 2008. "Knowledge and Action." *Journal of Philosophy* 105(10): 571-90.

Hieronymi, Pamela. 2006. "Controlling Attitudes." *Pacific Philosophical Quarterly* 87(1): 45-74.

Hyman, John. 2015. *Action, Knowledge, and Will.* Oxford: Oxford University Press.

Kelly, Thomas. 2011. "Following the Argument Where It Leads." *Philosophical Studies* 154(1): 105-24.

Marusic, Berislav, and John Schwenkler. 2018. "Intending Is Believing: A Defense of Strong Cognitivism." *Analytic Philosophy* 59(3): 309-40.

McDowell, John. 2001. "Might There Be External Reasons?" In *Mind, Value, and Reality.* Cambridge, MA: Harvard University Press.

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

Paul, L. A. 2015. Replies to Pettigrew, Barnes, and Campbell. *Philosophy and Phenomenological Research* 91(3): 794-813.

Paul, Sarah K. 2015a. Doxastic self-control. *American Philosophical Quarterly* 52(2): 145-58.

Paul, Sarah K. 2015b. 'The courage of conviction'. *Canadian Journal of Philosophy* 45(5-6): 647-69.

Van Fraassen, Bas. 2004. *The Empirical Stance*. Yale University Press.

Williamson, Timothy. 2000. *Knowledge and its Limits*. Oxford University Press.

ton Littlejohn, Eric Marcus, Laurie Paul, Sarah Paul, Richard Pettigrew, Juan Piñeros, Ted Poston, Jeremy Redmond, Jeff Sebo, and Marshall Thompson. Versions of it were presented at the 2015 Summer Seminar on the Nature and Value of Faith, the 2016 meeting of the Pacific APA, and workshops at Yale University and the University of North Carolina, Chapel Hill.

12. What Can Adaptive Preferences and Transformative Experiences Do for Each Other?[15]

Rosa Terlazzo

1. Introduction

In this chapter, I develop the helpful ties that exist between two concepts that have rarely been discussed together: adaptive preferences and transformative experiences. While each concept is rich and interesting in itself, each also raises unique problems that might be solved or allayed by appealing to and incorporating the other. But developing these ties also makes clear the limits of the work that they can do for one another—or at the least the limits of the work that can be done if we rely on a standard account of adaptive preferences. Ultimately, I argue that we can maintain the mutual fruitfulness of these ties if we develop an alternative account of adaptive preferences.

Adaptive preferences are, roughly, preferences that have been formed in relation to, and favor an option within, a limited option set. Standardly, they involve preferring less ambitious options that are available to more ambitious options that are not available, but would likely be preferred if they were. And indeed, in the cases of most concern to political philosophers, the adaptively preferred options are generally intuitively much worse than their unavailable counterparts. Think of the woman who endorses dangerous and unattainable beauty norms, the man who endorses rigid norms of masculinity that allow him little emotional expression or support, or the person who takes their spouse's controlling behavior to be welcome evidence of love. Since they involve a compromise for what can be had rather than a striving for what persons would otherwise seem to have strong reason to want, philosophers standardly take adaptive preferences to do a poor job of capturing persons' non-instrumental good. And insofar as they poorly capture persons' noninstrumental good in this manner, political philosophers

[15] Rosa Terlazzo, What Can Adaptive Preferences and Transformative Experiences Do for Each Other? In: Becoming Someone New: Essays on Transformative Experience, Choice, and Change. Edited by: Enoch Lambert and John Schwenkler, Oxford University Press (2020). © Rosa Terlazzo.
DOI: 10.1093/oso/9780198823735.003.00013

can appeal to the adaptiveness of adaptive preferences to explain why they should not play the same role that other preferences play in justifying public policy.

Transformative experiences, on the other hand, are experiences that change something significant about our relationship to the world.[1] Most importantly for our purposes, they can "change who you are, in the sense of radically changing your point of view" (Paul 2014: 16), "change your post-experience preferences, or change how your post-experience self values outcomes" (p. 48). In other words, they change your identity in a practical rather than metaphysical sense, by changing what you care about and how you orient your life. In this sense, experiences can be personally transformative. While the same experiences need not be personally transformative for everyone, parenthood, religious conversion, and adoption of a profession all serve as helpful illustrative examples of the kinds of experiences that frequently change our orientation towards the world in these fundamental ways. Less importantly for our purpose, experiences can also be epistemically transformative—and indeed, experiences which are personally transformative are often epistemically transformative as well. An epistemically transformative experience provides knowledge that one previously lacked, and that could not be acquired without undergoing the experience. That is, it generally provides "what it is like" knowledge that cannot be adequately acquired from other sources such as literature, scientific study, or testimony (Paul 2014: 10-11). So, since personally transformative experiences are often also epistemically transformative, they change us deeply in ways that we cannot in advance adequately anticipate.

While the concepts seem unrelated, the problems raised by each offer the possibility of fruitful interaction between the two. First, employing the standard account of adaptive preferences sometimes make it hard to show appropriate respect for persons with adaptive preferences, because it makes it hard to treat them as competent judges of their own good—at least with regards to their adaptive preferences. As I'll go on to argue, however, employing the concept of transformative experience can help us to explain how a person with an adaptive preference might eventually come to genuinely be benefited by the object of that preference. If this is right, then we can recognize that persons might both have adaptive preferences and be competent judges of their own good with regards to those preferences. Second, recognizing the reality of transformative experience raises problems for action guidance, since it is hard to know what to do when one lacks knowledge both of what an experience will be like and of whether to prioritize the preferences that she will have before the experience or the ones that she will have after. Happily, employing the standard account of adaptive preferences can give us at least some action guidance: it can tell us at least to avoid transformations that will cause us to develop adaptive preferences. But as I'll also go on to argue, it is hard to see how we can get both of these useful outcomes at once. Once we adopt the richer account of adaptive preferences that incorporating transformative experience al-

[1] Note that for the purposes of this chapter, I won't be distinguishing between transformative experience and nearby concepts like transformative choice. For this distinction, see Chang (2015).

lows, it becomes less clear why we should avoid developing adaptive preferences in the first place. And if we maintain the standard account of adaptive preferences that gives us clear guidance regarding transformative experience, it becomes harder to see how we can respect the relevant well-being judgments of persons with adaptive preferences.

I spend the bulk of the chapter developing the problems raised by using each concept to enrich the other. I end by briefly sketching the kind of alternative account of adaptive preferences that might allow us to escape this bind. While it joins the standard account in building a substantive evaluative judgment about prudential goodness into the concept of adaptive preference, it allows for the possibility that this judgment might cease to be applicable retrospectively. In this way, the alternative account retains a reason to avoid transformations that lead to adaptive preferences, while allowing that persons as they actually are post-transformation might be more significantly benefited by the objects of their adaptive preferences than by the objects that they would have preferred non-adaptively.

2. What Can Transformative Experiences Do for Adaptive Preferences?

Adaptive preferences raise significant problems in social and political philosophy. They are, roughly, preferences that have been formed in response to, and favor an option within, a limited set. The term "preference," however, tends to be used in a more general way by adaptive preference theorists than it is by decision theorists: while the latter take preferences to be only the ranking of one option relative to others in a specific set offered by a specific decision problem, the former use the term to refer to a less comparative attitude of general endorsement or acceptance. Yet even with this caveat, adaptive preferences at this broad level are relatively uninteresting. An overwhelming number of our day-to-day preferences are a result of exposure to some alternatives and lack of awareness of or exposure to others.[2]

The concept becomes more interesting when we consider preferences that are adopted as a way of reconciling ourselves to what is feasible for us or expected of us. Think here of the teenager who knows they cannot afford college and so decides that intellectual engagement is for nerds, or the adult who settles for the romantic partner who will take them despite disliking many of their traits. In such cases, getting something less feasible or less expected might be highly costly, and a failure to get it, or success in getting it that brings with it those great costs, might be deeply frustrating. So preferring what we take ourselves to be likely to get can be a way of

[2] Some moral philosophers are interested in the entirety of this broader set of adaptive preferences, and—unsurprisingly—argue that preferences do not warrant a negative evaluative judgment simply by virtue of being adaptive. Since these accounts are not included in the set of what I go on to call "standard accounts," I leave them to the side here. See Bruckner (2009); Dorsey (2017).

avoiding those costs while still keeping ourselves happy with the lives that we can have. Forming adaptive preferences, then, need not be irrational. We all face obstacles and opportunity costs in our lives, and when we simply cannot access some good—or when we must give up one legitimate good in order to access some other, greater good—we will often be genuinely better off if we can make ourselves happy with what we have. But note that saying that I am better off not wanting a thing that I cannot have is not the same as saying that the thing that I no longer want is not good for me. An object can be both highly valuable and something that I am better off not wanting, if wanting it when I cannot ever have it will only ever cause me frustration and unhappiness.[3]

This, then, is the idea captured by the standard account of adaptive preferences: they warrant a negative evaluative judgment that is not purely instrumental. That is, they involve the endorsement of an option which is in itself worse for us than some other object we could have preferred instead.[4]

And this standard account raises a problem for social and political philosophers: while it might be rational and/or desirable to form such preferences when the status quo seems unavoidable, those preferences can hardly be taken to justify the status quo. It is this point about justification that motivates social and political philosophers to adopt the use of the concept. As Mill rightly noted long ago, gender systems are as tenacious as they are in part because men have made women their allies in maintaining those systems (Mill 1989 [1869]). Indeed, one does not need to look far in contemporary society to find women who defend norms and institutions that keep them subordinate to men. Since a person's preferences about their own life generally carry normative force in determining the kind of society in which it would be good for that person to live, widespread acceptance of such norms and institutions among women should on the face of it give social and political philosophers reason to think that those structures and norms are legitimate. But the concept of adaptive preference offers them grounds on which to criticize such institutions and norms, even when those

[3] Plausibly, religions (such as Buddhism) and philosophies (such as Stoicism) that encourage us to avoid all attraction to worldly goods, do so because they simply deny that those goods are valuable in comparison to the enlightenment or detachment that one seeks to attain. I cannot address this substantive disagreement about value here.

[4] Note that this characterization of adaptive preferences is compatible with preferences formed either unintentionally or intentionally. While adaptive preferences are often assumed to develop "behind the back" of the person who has them (indeed, some philosophers even build this requirement directly into their account; see e.g. Colburn 2011), this characterization allows at least some instances of so-called "character planning" to also count as adaptive preference. (For this distinction, see Elster 1983.) In cases of character planning, one consciously and intentionally develops a new set of preferences with some goal in mind. That goal might well be to make oneself happy with what one is likely to get, but it might also be unrelated to constraining circumstances—e.g. one might engage in character planning in order to develop an appreciation for the finer things in life, or to heighten one's empathetic capacities. I will understand the former kind of character planning as an instance of consciously developed adaptive preference, and leave the latter to the side since it is irrelevant to our conversation.

disadvantaged by them seem to support them: If those disadvantaged persons endorsed the norms and structures only as a (conscious or unconscious) method of reconciling themselves to the limited and prudentially impoverished future they could expect to have, then this undermines the normative force of those preferences in supporting the status quo. The concept of adaptive preference, then, is generally employed by social and political philosophers as part of the expressly political project of criticizing marginalizing circumstances even when the marginalized have been made allies in the maintenance of their own plight.

Unfortunately, the same mechanisms that allow the standard account of adaptive preference to do useful political work also risk leaving the concept poorly placed to aid respectful treatment of those it is aimed to help. In the abstract, the concept is aimed at criticizing and thereby ending circumstances that marginalize vulnerable populations. But according to critics of the concept, the way in which it does so in practice risks further marginalizing them instead. Alison Jaggar, for instance, argues that appealing to adaptive preferences to cast doubt on social circumstances involves substituting the judgments of more privileged actors for the judgments of those who are already marginalized.[5] When disagreement arises, she takes appeals to adaptive preferences to be a way of ignoring or explicitly rejecting the judgments that marginalized persons make about their own lives. We might also worry that charges of adaptive preferences underestimate the rational and critical nature of the responses that marginalized persons have to their circumstances. H. E. Baber argues that using the concept discourages us from seeing those taken to have adaptive preferences as agents forced to make difficult choices without sufficient information or alternatives, and instead encourages us to see them as irrational, victimized, or psychologically damaged (2007: 126). And indeed, this tendency, as Uma Narayan points out, is tied up with a colonial history that encourages Westerners to "overemphasize constraints and underemphasize choice in other cultural contexts, while underestimating constraints and overemphasizing choice in Western contexts" (2002: 424). Like Baber, Narayan encourages us to see persons who seem to support the systems that marginalize them as actively bargaining with oppressive contexts to achieve the best outcomes that they can.

Ultimately, then, the general complaint is that appealing to adaptive preferences to criticize social circumstances involves making the judgment that marginalized persons who support those circumstances are mistaken about their own good—at least in that area. And this kind of judgement risks being seriously disrespectful of persons. Perhaps most obviously, it risks paternalistically denying them the opportunity to direct the course of their own lives—risks, in other words, a violation of what I have elsewhere called "primary recognition respect" (Terlazzo 2016). But even if persons with adaptive

[5] For something like this criticism, see Jaggar (2006: 317). Note that in this article, Jaggar does not direct this criticism at the concept of adaptive preferences directly, but rather at what she calls Nussbaum's "non- platonist substantive-good approach" to justifying capabilities. That approach, however, is motivated in large part by the problem of adaptive preferences.

preferences are allowed to direct their own lives—either because we trust their instrumental judgments about the good or because we are opposed to paternalism—such judgments risk the violation of a second kind of respect that I have elsewhere called "secondary recognition respect." Persons might endorse their purportedly adaptive preferences non-instrumentally, expressing them as values worth endorsing across a range of circumstances. And in these cases, a judgement that their preference is adaptive risks denying them equal moral status as persons competent to recognize their own good. Even if no risk of coercion is involved, this kind of disrespect remains worth taking seriously.

This is where the concept of transformative experience can prove helpful. If we take transformative experience seriously, then these judgments about the failures of persons with adaptive preferences to recognize their own good may not be warranted. Personal transformation, remember, changes our values, core preferences, and even characters in fundamental ways. And epistemic transformation renders us able to judge the value of an experience only after we ourselves have had it. These two things taken together might be sufficient to vindicate the self-regarding well-being judgments of those with adaptive preferences—and if they are, then they thereby block the grounds for disrespectful treatment that the concept seemed to recommend. The vindication would work in the following way. First, we would need to establish that different objects can be prudentially good for different people.[6] Second, we would need to establish that the kind of personal transformation that stems from transformative experience is sufficient to change persons in the ways that determine which objects are prudentially good for them. And third, we would need to establish that the development of an adaptive preference can involve changes that are sufficiently deep and fundamental to count as instances of personal transformation in the sense required by transformative experience. The concept of personal transformation clearly does most of the work here, since it is the possibility of personally transformative adaptive preferences that would allow the objects of those adaptive preferences to become newly good for their holders. But the concept of epistemic transformation also does important supporting work, since it explains why endorsement of adaptive preferences by their holders might be surprising to those without the adaptive preference in question. That is, if those without the adaptive preference have not undergone a relevant epistemic transformation, then they will be in no position to first-personally judge the well-being effects of that transformation.[7]

[6] One might think that this would seem controversial to objective list theorists, insofar as they hold that the same set of objective goods benefits everyone prudentially. This, however, gets things wrong: even objective list theorists who hold that the same type of good (achievement, wisdom, etc.) benefits everyone will generally recognize that different tokens of those goods (scientific achievement, athletic achievement, etc.) will be better for different persons.

[7] Note that personal transformation can occur in cases beyond those picked out by Paul's technical use of the category of transformative experience. While it is personal transformation generally that does most of the work here, I appeal to the narrower idea of transformative experience for two reasons.

In this chapter, I do not aim to establish any of these three claims (although I do so elsewhere[8]). Instead, I only aim here to give the outlines of the kind of plausible account we might offer if we wanted to vindicate the self-regarding well-being judgments of those with adaptive preferences. And if we can establish these three claims, then I think that we have such an account. If the three claims hold, then those with adaptive preferences might accurately judge themselves to be benefited by circumstances that they would not have endorsed before they developed their adaptive preferences. In other words, persons with adaptive preferences might have become persons who are genuinely benefited by their circumstances even if they did not begin as persons who were so benefited. In this way, we can recognize that persons with adaptive preferences might be right about their own good, and therefore have reason to take those with adaptive preferences seriously as both moral reasoners and agents competent and entitled to direct the course of their own lives. While it still remains an open question whether persons should have developed adaptive preferences in the first place (a point to which I turn below), the richer account of adaptive preferences that the concept of transformative experiences allows would give us reason to show respect for the judgments of those who have already developed adaptive preferences.

3. What Can Adaptive Preferences Do for Transformative Experience?

Transformative experiences raise at least two problems. First, insofar as they are epistemically transformative, we cannot know in advance what prudential value they will have for us. This makes it very difficult for us to make an informed decision in advance about whether it will be good for us to have them. The second is somewhat more complicated. The problem here is that insofar as they are personally transformative, it is possible that different things may be good for me before and after the experience. L. A. Paul (2014) offers the striking example of becoming a vampire. While I am still human, I may deeply value days spent in the sun, or eating rich and complex meals. But after I become a vampire, those pleasures might pale in comparison to the new ones available to me: swooping through the night, heightened senses, or whatever the case may be. If we allow that the things that I care about have a role to play in determining what is good for me, then it is once again difficult to make a choice about whether to undergo the experience, but this time for a different reason. The things that I value before and after the experience are incommensurable—I want different things

First, for the supporting role that epistemic transformation plays, discussed above. And second, because a focus on transformative experience highlights the extent to which the personal transformations we undergo are often a contingent result of the circumstances that life places us in, rather than of choices we freely make. Since adaptive preferences are also often the result of limited circumstances pushed on us by life, this is worth bearing in mind.

[8] Terlazzo (2017).

at different points, and we have no common scale of value that we can use to decide which is best. So, even if I knew in advance how I would be personally transformed, I would be stuck asking which set of goods I should prioritize in my decision: The ones I know I will come to value if I make the change? Or the different set that I in fact value now?

The problems raised are both problems about action guidance. What should we do, given that we cannot be certain how we will feel about the options that we have not yet experienced—and given that even if we did, we might still not know which set of incommensurable preferences to privilege? These are first of all questions that we must ask about our own choices. While self-regarding prudential considerations are rarely the only ones that we take to be relevant to our decision-making, they should play a significant role for those of us who care about promoting our well-being over the course of our lives. But many of us will also need to ask a second version of these questions. For those who have children or other vulnerable wards, as well as for those who otherwise have responsibility for influencing the course of vulnerable parties' lives, we also require action guidance regarding the transformative experiences that we should enable those vulnerable parties to undergo.[9]

Paul offers us some action guidance in answer to the first question. Even if we cannot rationally decide whether to have a transformative experience on the basis of what that experience will be like, we can still decide on other grounds that seem appropriately personal: We can decide on the basis of our desires to have a new experience for the sake of finding out what it will be like to be transformed (Paul 2014: ch. 4). Similarly, while it may not be possible to rationally decide who to become by comparing the value of having and satisfying the preferences we will come to have with the value of having and satisfying the preferences that we have now, we can still decide whether to be transformed for the sake of the transformation itself.

However, while this action guidance is helpful in some cases, it is decidedly less helpful in others. When deciding whether to have a child for the first time, it seems plausible that we could be reasonable to decide on the basis of whether we want to remain who we are or become a new person. The value of transformation and exploration themselves can plausibly tell us what to do. But it is less clear what action guidance we get in cases in which we must decide between alternative transformations in addition to deciding whether to be transformed at all. And many more of the decisions we make in life will be of this kind. The difficult question is not deciding whether to have a career, but which career to have; the difficult question is (often) not whether to move away from your home town, but where to move instead; and the

[9] I will refer in the remainder of the chapter simply to children, leaving out other vulnerable wards. This is in part but not only for the sake of simplicity. Since it is generally presumed that children will eventually develop capacities that make them answerable for themselves and their own decisions, a different set of issues arises for those (including the very old and those with serious cognitive disabilities or mental illnesses) of whom this cannot be said. However, doing justice to these differences is a large project beyond the scope of this chapter.

decision of whether to commit to a romantic partner is, for many of us, not a question of whether ever to commit at all, but whether to do it now with this person or later with another one. And when we must decide between a variety of different alternative transformations, deciding whether we value transforming at all will often be only the first and smallest step. We need, then, further action guidance.

Dana Howard (2015) focuses on offering us action guidance in regards to the second question: How should we decide for other vulnerable parties which transformative experiences they ought to undergo? Although the prudential considerations relevant to Paul's question certainly play a central and important role in answering Howard's question, the responses to the different questions ought to take different approaches to the maximization of interests. In cases in which the interests of others are not obviously at stake, it seems permissible for me to make my decisions only with my own prudential interests in mind. And while I plausibly have reasons to maximize my own utility, I do not seem to do anything morally wrong if I choose some sub-optimal course of action—or even some prudentially pretty miserable action—instead. But Howard's question concerns cases in which the interests of others are necessarily at stake. When an adult decides for a child, the interests of the vulnerable child will be constraining, since it is clearly morally impermissible to make decisions with prudentially miserable results for vulnerable others (at least assuming that better options are available). However, the child's interests will not solely determine the outcome, since the interests of the adults making the decision are also at stake. While parents certainly have reasons to maximize the expected utility of their children, increasing the utility of children will often have very real costs for their parents, and the interests of parents should also receive some weight. For instance, if moving to the other side of the world will be very slightly better for a child than remaining where she is, but extremely costly for the parent, then the parent does not seem to have an obligation to make that move—even if she has some reason to do so. Accordingly, we have some reason to narrow the focus of the second question: Which transformative experiences is it *permissible* for parents and other adults to facilitate children in undergoing?[10] This allows us space to consider parents' interests alongside children's, but also places some limits on the treatment that children may receive.

Indeed, this is precisely what Howard does. And here she holds that the concept of adaptive preferences can help to give us action guidance regarding the permissibility of causing a child to undergo a particular transformative experience: While the fact that the child will be glad in retrospect to have undergone an experience is some evidence that causing them to undergo that experience would be permissible, that evidence is overridden in cases in which the gladness seems to be the result of an adaptive preference.[11] Howard understands adaptive preferences as "preferences that

[10] Like me, Howard is interested in what is permissible *with regards to a child's own well-being,* and recognizes that other considerations may make a particular transformation all-things-considered permissible even if it significantly diminishes a particular child's well-being.

[11] I say "some evidence" because Howard holds that such gladness in fact only counts as evidence

are subconsciously formed in response to a person's diminished set of feasible options [... where] that person's preferences change to the point where the person prefers something that is within the feasible set of options" (2015: 361-2). What seems objectionable to Howard about preferences of this type is that "when this happens, the person's preferences become indistinguishable from accepting a sub- optimal situation" (p. 362). And what in turn makes it objectionable to cause transformations in children that lead them to have a diminished set of feasible life options—even when they are in retrospect glad to be in that position—is that we cannot reliably distinguish between cases in which that gladness is merely a successful attempt to make oneself content with a sub-optimal state and cases in which the gladness is more robust. Howard illustrates this point with the example of a famous clowning family. In the example, a father wonders whether to break his son's legs in order to cause the son to develop a ridiculous walk that will make him a better clown. Since his own father did so for him, and since he loves clowning and is glad that his walk enables him to do it so well, he reasons that his son is ultimately likely to be glad to have had his legs broken as well. Howard argues that even if the father's predictions are accurate, doing so would be impermissible. After all, "the whole point of breaking [the] son's legs is to diminish [his] feasibility set" so that he will become a clown, and if the son becomes glad to be such a good clown under these circumstances, we do not know whether his gladness is simply the result of coping with his diminished feasible set (2015: 363-4).

This action guidance has several attractive features. First, it allows us to continue to give weight to the child's own future happiness and the extent to which she will endorse her own life. In this way, it keeps her own judgments and assessment of her life front and center. But second, it also aims to ensure that our children will not be sold short by being forced to endorse less than they ought to. How far, though, does this action guidance take us? Remember, it tells us one set of conditions under which it is impermissible to cause a child to undergo a transformative experience, and it does so on the basis of her well-being—but it does not tell us which course of action it uniquely recommends pursuing. So while it tells us what not to do, it does not yet tell us what we ought to do. However, if the action guidance is appropriate (another point to which I return below), then this will still be quite helpful for parents deciding how to transform children, since it will counsel against a great many transformations that they might consider—namely, those that involve limiting a child's feasible set of options. So in answer to Howard's question regarding transformative experience, it seems that appealing to adaptive preferences provides us with some helpful action guidance.

Now consider Paul's question about transformative experiences: What about action guidance regarding the transformative experiences that I should choose to undergo for myself? Here the concept of adaptive preference might once again do some work.

---

when compared against the reactions the child would have to all of the possible alternative transformations (2015: 367). Since this distinction makes no difference for our purposes, I leave it to the side.

We might think that it cannot, since, as we saw, Howard appeals to adaptive preferences only to rule out some options, not to tell us which option among several we ought in fact to pick. And since Paul's puzzle begins with the question of how to maximize our expected utility when we necessarily lack information about what the utility of our options will be for us, we may be disappointed that we are once again left without guidance for deciding between options that are not ruled out. But while the action guidance that relies on adaptive preferences indeed does not tell us how to maximize our expected utility, it might still add something to our actionguidance toolbox. Remember Paul's own strategy. In absence of information about the utility of our post-transformation options, she changed the subject so that she now focused on a different but still appropriately personal consideration to which we could appeal. Rather than asking whether an experience whose contours we did not know would be enjoyable for us, she asked whether we valued making a change for the sake of having an experience whose enjoyment-worthiness we would go on to discover. But as we saw, while valuing the discovery that comes with making a change could help us to decide whether to make a change, it was not helpful when it came to deciding which change to make. Similarly, even if appeals to adaptive preference cannot tell us which option is best, they might still be able to tell us which options to avoid. After all, settling for the sub-optimal seems like something that each of us has prudential reason to avoid, so this seems like the right kind of person-considering reason to take into account.[12,13] And this might in turn help us to decide between the kinds of transformations that we might choose to undergo for the sake of discovery: all other things equal, we should avoid those that limit our feasible options, in order to reduce the chance that our future predicted endorsement will merely be the result of settling for the sub-optimal.

If all of this is correct, then the standard account of adaptive preference also has something to offer to work on transformative experience: The former concept can offer action guidance that helps us to narrow down the range of transformative experiences that we ought to undergo and encourage others to undergo.

[12] Remember that on Howard's account, having an adaptive preference does not necessarily involve settling for the sub-optimal. Rather, it opens up the possibility that we might be settling for the sub-optimal. In this way, then, her account is not one of the "standard" accounts to which I have been referring: i.e. it does not hold that the objects of adaptively formed preferences are (at least non-instrumentally) worse for us than the objects that we would have preferred non-adaptively. As we will see, however, Howard's type of action guidance only remains helpful if we introduce some kind of not-purely-formal evaluative standard into our account of adaptive preferences. And if we move in this fashion back towards the standard account of adaptive preferences, then adaptive preferences give us at least this version of helpful action guidance.

[13] Note that avoiding settling for the sub-optimal does not settle the issue. Some persons are more riskseeking than others, so they may be happier to risk some chance of settling for the sub-optimal if the alterative pay-off is a very valuable one that cannot otherwise be reached. However, a situation in which ones settles for the sub-optimal still seems to be one that each of us has reasons to avoid, all

4. Can We Do Both at Once?

Since each concept seems to gain something from the other, there seems to be significant value to discussing them in tandem. At this point, however, we need to ask whether these contributions can be made simultaneously. On the face of it, it seems difficult. Consider the general drive of Section 3: Roughly, we appealed to what was sub-optimal (i.e. adaptive preferences) to tell us what we ought not do when it comes to deciding between alternative transformative experiences. Now consider the general drive of Section 2: Roughly, we appealed to the concept of transformative experience to explain why something that seemed to be sub-optimal (again, adaptive preferences) might in fact be rightly taken to track significant value. So, the problem—once again roughly—is this: If we use transformative experiences to get a richer account of adaptive preferences, then can the concept of adaptive preferences really still offer us action guidance in which transformative experiences to undergo? It looks difficult, since the same feature of adaptive preferences—i.e. the standard account's commitment to a negative prudential evaluative judgment—is relied on by the second project but called into question by the first. So where are we left? It might seem that we must simply choose to give up on one of these seemingly valuable projects—either to reject the richer account of adaptive preferences, or to give up on the action guidance they might bring. While I think that we can ultimately avoid this seemingly forced choice by retaining but modifying the commitment to negative evaluative judgments found in the standard account of adaptive preferences, I turn now to explaining the apparent dilemma in greater detail.

To understand why we might be pushed towards choosing, let's return to one of the questions that I noted above required more attention: Is the action guidance that Howard offers really appropriate, on the rough definition of adaptive preferences that takes them to be those adjusted to a feasible set? The action guidance is supposed to be warranted because we introduce the possibility of sub-optimality when we form a preference in response to a limited set of feasible options. But is this really plausible? And if it is not, does the action guidance succeed? I'll argue that it is not simply the formally defined narrowing of option sets that introduces the possibility of suboptimality, and that the action guidance to simply avoid transformations that lead to a narrower option set therefore does not succeed. Instead, for the action guidance to be plausible, we must reintroduce a substantive evaluative prudential judgment into our account of adaptive preferences.

To show why the action guidance is not plausible as offered, we need to consider two questions. First, what should we take to narrow a set of feasible options?

Although Howard does not address this point directly, she raises it when comparing the case of a clown father to another case in which she thinks that a parent is not justified in transforming her child. In that case, a mother must decide whether to

---

other things being equal.

choose cochlear implants for her congenitally deaf infant. In making her choice, the mother recognizes that her child's life might be made much harder if he grows up deaf in a context like our own, which does little to accommodate deafness. But Howard denies that this recognition justifies giving the child cochlear implants, because, according to her, "these hardships don't necessarily diminish one's feasibility set, although they do make some goals harder to reach" (2015: 363). But what can we make of this distinction between difficulty and feasibility? And what is meant to rule an option infeasible if not difficulty?

If feasibility is understood in a binary sense, such that an option must be strictly inaccessible in order to be ruled out, then only the most extreme interventions will limit feasibility. Indeed, even the son whose legs are broken will not have his feasible set limited in an important way, since many people with non-standard bodies end up pursuing a variety of different life paths. This includes those with significantly less mobility in their legs than the son would have. Like the deaf child in a world not set up to accommodate deafness, the child with the broken legs will still have alternatives to clowning—they will simply be harder to access. But if feasibility is instead understood—as I believe we should understand it—as a matter of degree, then we must understand an option set to be narrowed either when options are removed entirely or when they are made more difficult to access. Otherwise, our definition will not capture the vast majority of intuitively adaptive preferences. Imagine a low- income student with an adaptive preference to forgo college, or a woman in a very patriarchal religious community with an adaptive preference to forswear a life of independence outside of it. In each case, the options not preferred could be accessed, as evidenced by the few similarly placed persons who manage to access them. The fact that more similarly placed individuals do not also do so is evidence that doing so is genuinely hard. This difficulty is a real limit to their feasible sets of options, despite the fact that it does not entirely rule out the other options. So if Howard defends the decision to forgo cochlear implants by appealing to the distinction between difficulty and feasibility, then the concept of feasibility no longer helpfully illuminates the concept of adaptive preferences.

Howard does, however, offer a secondary justification to treat the cochlear implant case differently, which might prove more helpful: She argues that "when the hardships that are presented to a deaf person become diminishments of that person's feasibility set, this is not necessarily a result of their impairment, but rather a result of the way that their impairment is accommodated by society" (2015: 363, my emphasis). Therefore, opting for cochlear implants on the basis of those hardships "means accepting unjust conditions of society as fixed features of the condition of the disability" (p. 363). This justification, rather than holding that the child's feasible set will not be narrowed, holds that we have political obligations to fight injustice rather than concede to it. It is surely true that we have political obligations to fight injustice, and it may well be the case that rejecting cochlear implants is what is called for as a way of fighting injustice. But note that in this case, the action guidance regarding the transformation is not

provided by the concept of adaptive preferences, and even conflicts with the action guidance that the concept was otherwise supposed to offer. As long as the parent reasonably believes that the child's life will be made significantly harder by being deaf, and that this difficulty will limit her option set in the actual world, then she should also recognize that her child's eventual endorsement of being deaf may well be the result of an adaptive preference. If Howard is still right about the permissibility of forgoing cochlear implants, then she must either justify that claim despite the possible adaptiveness of the preference that will arise, or else adopt a different and richer account of adaptive preferences.

The second question that this action guidance requires us to ask ourselves is whether it is always bad to have our set of feasible options narrowed. Some feasible options seem to be quite bad for us upon consideration, even while appearing attractive initially. With regards to these options, our lives might go uncontrover- sially better if we did not have them, and were therefore not tempted to pursue them. Indeed, many of the public health measures that tend to make us healthier and happier are justified in this way: By removing superficially attractive options that are bad for us, or making them harder to take advantage of, these measures make us better off. In the same way, I might narrow my child's option set so that taking all very seriously addictive drugs is off the table. Many of us will hold that my child's life is not worse for this, despite the fact that I have narrowed her set of options.[14] So why are we justified in thinking that it is necessarily bad to have our option sets narrowed?

Note that option sets might be narrowed in at least two ways. First, an option set might become narrower if you remove one qualitatively different option from the person's range, as in the case of removing the option for using hard drugs. But an option set might also be narrowed if a new, narrower range of options is swapped for an older, different, and broader range. Howard might object to my point about hard drugs by noting that she is instead concerned with the second case—the case in which a person ends up with a very different and much narrower range of options. This alternative is suggested when she claims that the mother could permissibly decide to forgo cochlear implants because she could reasonably hold that her baby "can remain profoundly deaf and still pursue the same variety of worthwhile life plans as someone with cochlear implants" (2015: 364). This is obviously false on one reading: The substantive set of life plans open to each possible child will be very different. But on another reading, it may be true: The children may have access to a comparably broad range of valuable life plans, although these plans will be different. If this is what Howard intends, then my example of addictive drugs might not make trouble, since we seem to be removing only life plans that are not worthwhile. Perhaps that is so, and what matters is that a child is not being given a narrower set of valuable life plans than she would otherwise have. But note that we are not then making a formal claim about narrower and wider

[14] At least, it does not seem to be worse from the point of view of well-being. Those committed to certain very strong accounts of autonomy, say, may still hold that it is worse on some other dimension.

options sets. Instead, we are making a substantive claim about which options count as worthwhile. And once we are making claims about which lives are worthwhile, then we should presumably care about the degree of value, and count that into the evaluation of an option set. In that case, a large range of relatively valuable possible lives might be worse than a narrower range of much more valuable ones. And if this is right, then evaluating the decision of the clown father is much more difficult. Now we cannot simply appeal to the fact that he narrows his son's option set by breaking his legs. Instead, we are now involved in a substantive debate about the comparative value of a life of clowning and all other lives. If we want to go down this road, it seems that we will need to reintroduce a substantive evaluative judgment in order to determine which adaptive preferences ought to be avoided. The simple, formal account of adaptive preferences, then, does not seem to give us much action guidance regarding the transformative experiences that we may permissibly cause our children to undergo.

While I cannot in this chapter canvass every possible way of reincorporating substantive evaluative judgments, let's consider the prospects for reintroducing the substantive evaluative judgement offered by standard accounts of adaptive preferences: That they necessarily involve the endorsement of an option that remains in itself worse for us than the object of some other preference we might have otherwise had. We can take Serene Khader's (2011) prominent perfectionist account as one such example.[15] For Khader, "adaptive preferences are (1) preferences inconsistent with basic flourishing (2) that are formed under conditions non- conducive to basic flourishing and (3) that we believe people might be persuaded to transform under normative scrutiny of their preferences and exposure to conditions more conducive to flourishing" (2011: 42). This sort of perfectionist account would give us the tools to distinguish between the permissibility of breaking legs and forgoing cochlear implants. We would simply need to establish that breaking the son's legs would be inconsistent with and would therefore compromise his basic flourishing, while growing up deaf would not be inconsistent with and therefore would not compromise the basic flourishing of the other child. That account, then, would provide us with action guidance regarding the two cases.

There is, however, a significant cost to this action guidance: It is hard to see how the concept of transformative experience can then do the work for the concept of adaptive preference that we wanted it to do in Section 2. In other words, it is hard to see how we could show respect for persons who had and endorsed adaptive preferences. Remember the work that transformative experiences were supposed to do. They were supposed to explain how a person could be transformed by the experience of developing her adaptive preference such that the object of her preference was now one that had genuine value for her. But on Khader's perfectionist account of adaptive preferences,

[15] Khader's account of course represents only one way of incorporating a substantive evaluative judgment—indeed, I myself recommend a different way of incorporating a substantive evaluative judgment towards the end of the chapter. However, the criticisms of Khader's account, suitably modified, should apply to any account that takes the substantive judgment to apply both before and after the preference has been developed.

we cannot tell this kind of story. Insofar as an adaptive preference is by definition one that is inconsistent with basic flourishing, it would be inappropriate to hold that the person's adaptive judgment about her own good could be correct. And insofar as the development of the adaptive preference genuinely transforms her such that she comes to prudentially benefit from it, it ceases to be adaptive. In other words, if our definition of adaptive preferences incorporates fixed perfectionist ideals of human flourishing, then we cannot appeal to transformative experiences to explain why we should trust some of the genuinely adaptive judgments of those with adaptive preferences. In this way, we provide action guidance for transformative experiences, but lose our justification for showing persons with adaptive preferences the particular kind of respect that an appeal to transformative experiences was meant to provide.[16]

If respect is important, then one option is to simply give up on the work that adaptive preferences can do for transformative experiences, and try to maintain the beneficial relationship between the concepts only in the other direction. If we give up on the perfectionist account of adaptive preferences and give up on action guidance for transformative experiences, then transformative experiences can still explain how and why individuals might rightly judge that their adaptive preferences correctly capture their own good. In that case, discussing the two concepts in tandem at least still gives us reasons to respect those with adaptive preferences.

To see why this route is not attractive, we must return to the point of discussing adaptive preferences in the first place. Remember that adaptive preferences were supposed to help us with a political project; they were supposed to explain why we ought to criticize marginalizing circumstances even when the marginalized endorsed them. But if the marginalized might have been transformed such that they are now right to endorse those circumstances, then what grounds do we have for criticizing them? In other words, if appealing to transformative experiences lets us hold that persons with adaptive preferences might rightly judge their own good even in the case of their adaptive preferences, then is there even a justification for using the concept of adaptive preferences in the first place?

At this point we seem to be in an unattractive bind. It seems that we have two choices: (i) succeed in showing this type of respect for persons with adaptive preferences, but give up on both the action guidance that adaptive preferences provide for transformative experiences and the political project that motivated the use of adaptive preferences in the first place, or (2) retain the action guidance and the motivation for

[16] Note that since Khader's account relies on the idea of basic flourishing, and stipulates that adaptive preferences are those that we should expect to revert or transform under conditions conducive to flourishing, the concept of adaptive preferences on her account gives us little action guidance with regards to transformative experiences. All action guidance work is done primarily by the directive to avoid transformations that compromise our basic flourishing, not to avoid adaptive preferences. And indeed, on her definition adaptive preferences seem to be if anything preferred among the set of transformations involving a deficit of basic flourishing, since we should expect them to revert under normative scrutiny and better conditions.

the political project, but give up on our richer and more respectful account of adaptive preferences.

5. Escaping the Bind

To get out of this bind, we can modify the standard account's negative evaluative judgment, rather than adopting or rejecting it wholesale. To see how we might do this, let's return to the other point that I noted above would require further discussion. I said that the richer (i.e. transformative-experience-incorporating) account of adaptive preferences would give us reason to show respect for those persons who have already developed adaptive preferences while leaving open the question of whether those persons ought to have prudentially developed those adaptive preferences in the first place. If the answer to this question is that they had some intrinsic prudential reason not to do so, then the concept of adaptive preferences will remain useful for the political project. In other words, we can try to establish that we can have robust prudential reason for satisfying an adaptive preference after it has been formed, despite having had prudential reason not to develop it in the first place.

Ultimately, for our twin projects to be successful, we will of course require a theoretical account of precisely what this prudential reason is. And what that reason looks like will depend on the account of prudential goodness—that is, the account of well-being—that we use. I am optimistic that we can give a successful account of such a reason. Indeed, I have elsewhere argued that all major theories of well-being have the resources to plausibly explain how one can have both a robust prudential reason to satisfy an already-formed adaptive preference and a robust prudential reason not to form that preference in the first place (Terlazzo 2017). The substance of that reason, however, will vary depending on the theory of well-being we use, and determining which theory we ought to use is itself a project too large for a single paper.

Since I cannot complete that project here, let me end with an example that both offers intuitive support for my optimism that a reason can be offered and gives us some sense of how that reason might work in practice. I modify the example from one offered by Elizabeth Harman (2009). In this example, we consider a young teenager who might go on to have two possible futures. In one, she gives birth to a child at 15, and goes on to develop a wonderful relationship with that child. While she goes on to have a good life, the difficulty of raising the child as a teenager means that she struggles, and her relationship with her child accordingly receives less nurturance than it could have. In the other, she has a child in her early 30s, with whom she also goes on to have a wonderful relationship. However, in this possible future, she spends her earlier years in many intrinsically valuable pursuits, and has more time and resources to devote to her child and their relationship by the time the child is born. In retrospect, both possible versions of the mother clearly have intrinsic reason to be glad they had the lives that

they in fact had—but prospectively, the teenager clearly has more intrinsic prudential reason to take the second path than the first.

This is the intuitive reason that must be given theoretical legs by our account of well-being. Much of what I have said here has focused on modifying standard accounts of adaptive preferences in this way in order to allow our two concepts to work fruitfully together. But note that in order for them to do so in the way in which I have proposed, the way in which we evaluate transformative experiences is also modified. On Paul's account of transformative experiences, the only values that we can appropriately appeal to in order to evaluate transformations are the agent's own (time-indexed) preferences. If those preferences are incommensurable, we are left without prudential guidance. But when we offer to our future mother the intuitive reason for choosing one course of action over the other, we do not appeal only to the preferences held by each version of herself. Instead, we appeal to a theory of wellbeing that can ground that reason. To be sure, that account of well-being will make concessions to the ideas of personal and epistemic transformation. It is personal transformation that allows each later version of the mother to be rightly glad that she has the life that she does. And it is epistemic transformation that makes each mother best placed to judge whether there are grounds to be glad. But by appealing to a theory of well-being before either transformation has been undergone, we can introduce a relevant value into the decision that goes beyond the values offered by the incommensurable preferences of her future selves.

6. Conclusions

Taken together, the dual temporal perspective on adaptive preferences and the modification to the information used to evaluate transformative experiences would allow the concepts of adaptive preferences and transformative experiences to simultaneously do for each other all three pieces of valuable work we have identified here. First, we can show respect for persons with adaptive preferences. Because adaptive preferences could accurately capture a person's good post-transformation, transformative experiences explain why we have reason to treat persons with adaptive preferences as authorities on their own good. Of course, like any other person, those with adaptive preferences might change their preferences as a result of deliberation and exposure to new alternatives, and we therefore have reasons to pursue this deliberation and exposure for those with adaptive preferences. But if their preferences do not in the end change, then the possibility of transformative experience will explain why we have reason to respect those preferences and treat individuals in accordance with them. In this way, the richer account of adaptive preferences that transformative experience gives us provides us with non-political reasons to show respect for persons with adaptive preferences. Given their transformation, their adaptive preferences now directly warrant being given weight in determining how to treat them.

Second, we can offer some action guidance for the transformative experiences we undergo. The same set of modifications allows us to say that the same adaptive preferences that go on to actually capture a person's good post-transformation are still ones that she lacks an important prudential reason to develop pretransformation. In the case of such preferences, we have legitimate prudential reason to aim to prevent persons from undergoing transformative experiences involving those adaptive preferences in the first place. To be sure, this action guidance is once again incomplete. To the extent that personally transformative experiences are also epistemically transformative, we may not be able to tell in advance the extent to which a transformation will benefit us, and this can make it hard to accurately identify the relevant set of adaptive preferences prospectively. But once we are aware that such adaptations are possible, we can at least look out for them, and attempt insofar as we can to identify and avoid them. We are at least no longer left with nothing to appeal to beyond two incommensurable sets of preferences.

Third and finally, these modifications allow the concept of adaptive preference to remain useful for the political project. Even if we cannot criticize social circumstances for currently harming persons who have already been transformed in accordance with them, we can criticize those circumstances for causing future people to undergo transformations that they would be better off not undergoing. While I anticipate that providing the full solution would require a book-length project, we should be hopeful that the concepts of transformative experience and adaptive preference can both simultaneously enrich and inform one another.

References

Baber, H. E. 2007. "Adaptive Preference." *Social Theory and Practice* 33(1): 105-26.

Bruckner, D. W. 2009. "In Defense of Adaptive Preferences." *Philosophical Studies* 142(3): 307-24.

Chang, R. 2015. "Transformative Choices." *Res Philosophica* 92(2): 237-82.

Colburn, B. 2011. "Autonomy and Adaptive Preferences." *Utilitas* 23(1): 52-71.

Dorsey, D. 2017. "Adaptive Preferences Are a Red Herring." *Journal of the American Philosophical Association* 3(4): 465-84.

Elster, J. 1983. *Sour Grapes: Studies in the Subversion of Rationality.* Cambridge: Cambridge University Press.

Harman, E. 2009. " 'I'll Be Glad I Did It': Reasoning and the Significance of Future Desires." In J. Hawthorne (ed.), *Ethics,* 177-99. New York: Wiley Periodicals.

Howard, D. S. 2015. 'Transforming Others: On the Limits of 'You'll Be Glad I Did It' Reasoning." *Res Philosophica* 92(2): 341-70.

Jaggar, A. 2006. "Reasoning About Well-Being: Nussbaum's Method of Justifying the Capabilities Approach." *Journal of Political Philosophy* 14(3): 301-22.

Khader, S. J. 2011. *Adaptive Preferences and Women's Empowerment.* Oxford: Oxford University Press.

Mill, J. S. 1989 [1869]. "The Subjection of Women." In *On Liberty and Other Writings,* ed. Stefan Collini. Cambridge: Cambridge University Press.

Narayan, U. 2002. "Minds of Their Own: Choices, Autonomy, Cultural Practices, and Other Women." In L. M. Antony and C. E. Witt (eds), *A Mind of One's Own: Feminist Essays on Reason and Objectivity*, 418-32. Boulder, CO: Westview Press.

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

Terlazzo, R. 2016. "Conceptualizing Adaptive Preferences Respectfully: An Indirectly Substantive Account." *Journal of Political Philosophy* 24(2): 206-26.

Terlazzo, R. 2017. "Must Adaptive Preferences Be Prudentially Bad for Us?" *Journal of the American Philosophical Association* 3(4): 412-29.

13. Punishment and Transformation[16]

Jennifer Lackey

1. Introduction

In this chapter, I argue that strict, long-term punishments are epistemically irrational. By appealing to the radical changes in mental states that can be brought about through transformations or transformative experiences, and showing that punishment needs to be sensitive to such mental states, I argue that strict, longterm punishments screen off the possibility of being sensitive to epistemic information that is highly relevant. I conclude that rationality demands that justice be an ongoing process, open to revision in light of changes in both those being punished and those doing the punishing.

2. Preliminary Remarks

It will be helpful to begin with some clarifying remarks about the terminology I will be using, as well as the general framework for the issues to be discussed. By "strict," I mean punishments that are not open to revision. Prison sentences without the possibility of parole are classic examples of strict punishments as I am understanding them here. Moreover, I am distinguishing between a punishment being revised and one being overturned. A prison sentence without the possibility of parole cannot be reduced or otherwise modified, but it can be overturned if, for instance, exculpatory evidence is discovered. "Long-term" will obviously be a somewhat loose notion, and may even depend in part on the nature of the action being punished. Losing the car for a year might be a long-term punishment for a teenager violating curfew, for instance, while this same amount of time might seem short-term for a violent offense. My purpose here is not to wade into questions of this sort. Instead, I will focus on clear, paradigmatic cases of long-term punishments and argue that they are epistem- ically irrational. To

[16] Jennifer Lackey, Punishment and Transformation In: Becoming Someone New: Essays on Transformative Experience, Choice, and Change. Edited by: Enoch Lambert and John Schwenkler, Oxford University Press (2020). © Jennifer Lackey.
DOI: 10.1093/oso/9780198823735.003.00014

the extent that the premises on which my arguments rely are true of other punishments, they can be applied accordingly.

I will often use natural life sentences as a central case of a strict, long-term punishment. A sentence of "natural life" means that the remainder of one's natural life will be spent behind bars. There are no parole hearings, no credit for time served, no possibility of release. Short of a successful appeal or an executive pardon, such a sentence means that the convicted will, in no uncertain terms, die behind bars.

While natural life sentences will be the central case of strict, long-term punishments, they are by no means the only one. De facto life sentences, which are sentences that exceed the life expectancy of the convicted, are also clear examples. A 25-year-old man who is given a 6o-year sentence, for instance, will most likely end up with the same outcome as a 25-year-old given a natural life sentence: death while in prison. Indeed, this is even clearer when we factor in the toll that incarceration takes on mortality rates. While the average life expectancy of an American is 78.8 years (Fast Stats 2016), it is 64 years for those who are incarcerated ("Michigan Life Expectancy Data For Youth Serving Natural Life Sentences," n.d.). Thus, even many significantly shorter sentences turn out to be de facto life ones once this is taken into account.

Moreover, truth-in-sentencing policies, which "require those convicted and sentenced to prison to serve at least 85 percent of their court-imposed sentence," turn many lengthy prison sentences into strict ones (Olson et al. 2009). For instance, the state of Illinois adopted in August of 1995 its version of truth-in-sentencing, which requires, among other things, that those convicted of murder serve 100% of their sentences (Olson et al. 2009). Given this, every sentence for murder in Illinois is a strict one, with no parole hearings and thus no possibility of revision.

The scope of this chapter will be the question of the rationality of given punishments, even if they have been appropriately applied at the outset. A variety of issues might make us rethink our punishment policies. For example, in recent years there have been many cases of people being exonerated after it has been discovered they were wrongfully convicted. There has also been an increasing appreciation for the ways in which structural racism and economic disadvantages affect how punishments are distributed. These are very serious concerns, but I will set them aside here to focus instead on another issue—the ongoing rationality, and therewith justice, of strict, long-term punishments, regardless of their origin.

Finally, all of my arguments will be epistemic in nature, with the conclusion that strict, long-term punishments are epistemically irrational. By "epistemic," I mean of, or related to, knowledge and knowledge-related goals. For instance, it might be rational from a moral point of view to believe in your son's innocence, despite the overwhelming evidence on behalf of his guilt, because of the moral duties parents have to their children. It might also be practically rational to believe that you're going to survive your terminal diagnosis in the face of massive amounts of evidence to the contrary, because this belief will significantly improve the quality of your remaining days. But believing against the evidence is clearly at odds with achieving knowledge- related

goals—such as truth, justification, understanding, and so on—and thus in neither case would the agent involved be epistemically rational.

3. The Transformation Argument

With these points in mind, let's turn to my first argument on behalf of the conclusion that strict, long-term punishments, such as natural life sentences, are epistemically irrational. I will call this the Transformation Argument (TA).

The first premise of the TA is:

> (i) Punishment at a given time ought to be sensitive to the relevant evidence available at that time.

One way of supporting (i) is by appealing to a widely accepted view in epistemology, according to which action is epistemically appropriate only when it is grounded in a sufficiently good epistemic position. A central candidate for this position is knowledge. For instance, Timothy Williamson maintains that the "epistemic standard of appropriateness" for practical reasoning can be stated as follows: "One knows q iff q is an appropriate premise for one's practical reasoning" (Williamson 2005: 231).[1] Similarly, according to John Hawthorne and Jason Stanley, "Where one's choice is p-dependent, it is appropriate to treat the proposition that p as a reason for acting iff you know that *p*" (Hawthorne and Stanley 2008: 57s).[2] We can call the thesis found in these passages the Knowledge Norm of Practical Reasoning, or the KNPR, and formulate it as follows:

> KNPR: It is epistemically appropriate for one to use the proposition that p in practical reasoning if and only if one knows that p.

As stated, there are two dimensions to the KNPR; one is a necessity claim and the other is a sufficiency claim. More precisely:

> KNPR-N: It is epistemically appropriate for one to use the proposition that p in practical reasoning only if one knows that p.
>
> KNPR-S: It is epistemically appropriate for one to use the proposition that p in practical reasoning if one knows that p.

[1] This is the insensitive invariantist version of the KNPR. The contextualist version is: "A first-person present-tense ascription of 'know' with respect to a proposition is true in a context iff that proposition is an appropriate premise for practical reasoning in that context" (Williamson 2005: 227).

[2] Hawthorne and Stanley restrict their conditions to "p-dependent choices" since p may simply be irrelevant to a given action. Since the cases discussed here all involve p-dependent choices, I will ignore this complication in what follows.

While both versions of the Knowledge Norm of Practical Reasoning are of interest, my focus here will be on only the necessity claim.

Actions are typically the result of practical reasoning, and those actions that are epistemically proper must be the product of practical reasoning that itself meets a sufficient epistemic standard.[3] If, for instance, the KNPR-N is correct and this standard is knowledge, then acting on p requires that one know that p.

In place of KNPR generally and KNPR-N more specifically, others have argued for weaker constraints on practical rationality and action. Ram Neta, for instance proposes the following:

> JBK-Reasons Principle: Where S's choice is p-dependent, it is rationally permissible for S to treat the proposition that p as a reason for acting if and only if S justifiably believes that she knows that p. (Neta 2009: 686)

On Neta's view, then, epistemically proper action requires, not an appropriate connection with knowledge, but, rather, justified belief that one knows. In particular, in order for it to be epistemically proper to act on p, one must justifiably believe that one knows that p. Others have defended even weaker views, especially with respect to particular kinds of action,[4] such as a *justified belief* norm (Kvanvig 2009),[5] a *rational credibility* norm (Douven 2006), a *reasonable-to-believe* norm (Lackey 2007), and a *supportive reasons* norm (McKinnon 2013).

It is not my aim here to defend one of these views over another. Instead, my point is to make clear that it is widely accepted that there *is* an epistemic standard governing action, regardless of the strength of the norm. Support for this more general point comes from our practices of praise and blame. If, for instance, you fail to pick up my daughter from school today at 3:35 p.m., despite the fact that you do so every weekday, an appropriate defense would be, "I knew, or I had good reason to believe, that all students needed to stay until 4:00 p.m. today because this is what the school's website indicated." Your action here would be justified by appealing to the epistemic status of the beliefs guiding it and, assuming that they meet the correct standard, you wouldn't be subject to criticism for failing to pick up my daughter.

This brings us back to premise (1). Punishment is an action, and so in order for it to be epistemically proper, it must meet an epistemic standard. Regardless of whether this standard is knowledge or something weaker, none permit ignoring relevant evidence. Indeed, even if knowledge is understood in a highly externalist way, such that it requires something along the lines of reliability or truth-tracking, evidence cannot be disregarded. Counterevidence, for instance, can always function as a defeater, ei-

[3] Even if an action is not actually the product of practical reasoning, it might still be the case that it needs to be capable of being endorsed by practical reasoning, if one were to engage in it.

[4] This is prevalent in the literature on the norm governing epistemically proper assertion.

[5] It should be noted that this norm is distinct from Neta's *justified-belief-that-one-knows* norm.

ther a rebutting or an undercutting one.[6] Even if, say, I reliably form the belief that there is a fox in your backyard on the basis of perception, evidence that my belief is false (rebutting), or unreliably formed or sustained (undercutting), might still defeat my justification. If you tell me that foxes have never been seen in your area, and that your neighbor has a Shiba Inu who frequently escapes into your yard, then I have evidence against the truth of my fox belief. Similarly, if my optometrist tells me that I'm wearing glasses with a wildly incorrect prescription, then I have evidence against the reliability of the basis of my fox belief. Even if we are not all evidentialists, according to which evidence is the central source of epistemic goods, it is clear that there is no way of circumventing the crucial role that evidence plays in our beliefforming practices. As David Hume famously said, "a wise man proportions his belief to the evidence."

Moreover, relevant evidence includes not only what is possessed at a given time, but also what a subject *should have* at that time. Suppose, for instance, that the total body of evidence that a racist possesses justifies his white supremacist beliefs, but only because he deliberately avoids the acquisition of any evidence to the contrary.

Perhaps he carefully chooses to be surrounded by only other racists, and he reads only news sources that defend his beliefs. Even though the evidence in his possession supports his racist beliefs, it doesn't follow from this that they are justified. Why not? Because there is evidence that he ought to possess, or ought to consider. Such evidence is often characterized in terms of normative defeat, in contrast to doxastic *defeat,* which involves counterevidence that is already possessed.[7] But the central point is that "relevant evidence" in (i) is not limited to the evidence that a subject possesses. This is important, especially in matters of criminal justice. It clearly won't do for a police officer to fail to follow through with highly relevant and credible leads, only to fall back on the view that his belief in the defendant's guilt is justified relative to the wildly incomplete evidence he in fact possesses. Similarly, it won't do for the state to ignore who prisoners have become after serving decades in prison, only to claim that the very lengthy sentences are justified relative to the evidence that was had at the time of their convictions.

Support for (i) is thus found in the combination of the following three claims: (i) action is governed by an epistemic norm; (ii) any version of this norm will include room for relevant evidence; and (iii) punishment is an action. Otherwise put, punishment, being an action, should be grounded in epistemically proper beliefs of the one doing the punishing, such as the state, and epistemically proper beliefs cannot be insensitive to relevant evidence. Thus, punishment at a given time ought to be sensitive to the relevant evidence available at that time.

---

[6] For various views of defeaters, see BonJour (1980; 1985); Nozick (1981); Goldman (1986); Pollock (1986); Fricker (1987; 1994), Chisholm (1989); Burge (1993; 1997); Plantinga (1993); McDowell (1994); Audi (1997; 1998); Bergmann (1997); Williams (1999); BonJour and Sosa (2003); Hawthorne (2004); Reed (2006); and Lackey (2008).

[7] See Lackey (2008). See also Goldberg (2017) for a very illuminating account of when a subject "should have known" a given proposition.

Such a view is certainly borne out in many of our practices. Perhaps the most striking is found in exonerations, many of which are made possible by evidence uncovered even decades after the original convictions. Since 1989, for instance, there have been 367 people in the USA exonerated by DNA testing ("DNA Exonerations in the United States" n.d.). In many of these cases, it is only because of the technological evolution of forensic DNA profiling that evidence became available to prove the innocence of the defendants. Were future evidence screened off so that punishment was based only on the evidence available at the time of the original convictions, the innocence of these hundreds of men and women would never have been established. Similar considerations apply to other sources of exoneration—a key eyewitness recants, testimony becomes available that the defendant was tortured and issued a false confession as a result, theories once accepted by the scientific community become widely challenged, such as evidence of arson or Shaken Baby Syndrome,[8] and so on. In each of these cases, our practices in the criminal justice system clearly support (1). Indeed, were this not to be the case—were our criminal justice practices to be at odds with (1)—epistemic and moral errors of massive proportions would be rampant.

The second premise of the Transformation Argument is:

> (2) Punishment ought to be sensitive to the mental states of the one being punished, where this includes his or her mental states after the time of the punishable act.

By "mental states" here, I'm including beliefs, desires, emotions, intentions, and dispositions. Mental states that are paradigmatically relevant are, for instance, whether the punishee appreciates the wrongness of his or her actions, feels remorse, intends to avoid wrongdoing in the future, and so on. Moreover, notice how weak this premise it. It does not say that punishment ought to be sensitive to only the mental states of the punishee. Thus, many other factors might bear on punishment, such as retribution, deterrence, and the mental states of the victims, if there are any. It also does not require that punishment be highly sensitive to the punishee's mental states. In this way, punishment may even be largely determined by factors besides the beliefs, desires, emotions, intentions, and dispositions of the one being punished. Perhaps the primary aim of punishment is to serve as a deterrent, and the mental states of the punishee are relevant only insofar as they bear on or supplement this. But what this premise does make clear is that the mental states of the person being punished cannot be screened off when considering the legitimacy of a given punishment.

Support for this premise can be found in many of our practices. The most obvious is that the mental states of a person at the time she commits a punishable act are often taken to be of great significance when determining the punishment deserved. A clear example here is the role that mens rea, or a "guilty mind," plays in the criminal justice system. For instance, a person who engages in illegal activity because he or she

[8] See e.g. Tuerkheimer (2014).

honestly misperceives reality—that is, he or she makes a "mistake of fact"—lacks mens rea and is said to be such that he or she ought not be charged with, or convicted of, a crime. If, say, I reasonably but mistakenly believe that your MacBook Air is mine and I leave our department meeting with it, the absence of a guilty mind means that I shouldn't be charged with, or convicted of, theft. Similar considerations apply to the difference between first-degree murder and lesser charges, such as second-degree murder or manslaughter: only the former requires premeditation which, in turn, involves mental states, such as the intention to kill, beliefs about how to bring this about, and so on. Indeed, some convictions, such as those involving conspiracy charges, are grounded entirely in the mental states of the accused. Clearly, then, the very nature of the criminal charges brought against people depends heavily on their mental states, thereby directly affecting the punishment, or lack thereof, handed out to them.

Even more importantly for the purposes of this chapter, however, is that we often regard the mental states of a punishee long after a punishable act has been committed to be relevant to the punishment deserved. In explaining given sentences, for instance, judges often appeal to the mental states of the convicted at the time of the trial or the sentencing, even if this is long after the crime in question was committed. This is clear in the recent highly controversial case of Brock Turner, the student from Stanford who was convicted of three felony counts of sexual assault and sentenced to six months in county jail, three years of probation, and a requirement that he register as a sex offender. In defending what many regarded as an egregiously light sentence, the judge "said that he believed the defendant felt genuine remorse" (Svrluga 2016). Now it is clear that the judge is not speaking about Turner's mental states at the time of the sexual assault, nor even immediately afterward, but, rather, at the time of the trial and sentencing, which occurred well after a year of the crimes themselves. Moreover, even though there was a public outcry about the punishment Turner received, many saying that it was disproportionate to the seriousness of the crimes, the criticism did not focus on whether Turner's remorse, or lack thereof, ought to have made a difference. Indeed, when Turner himself seemed to blame what occurred on a "party culture" of "drinking," the fact that he seemed to fail to appreciate his own agency in the sexual assault that he perpetrated fueled calls for a more stringent sentence (Levin and Wong 2016). Thus, even though the judge and the public are at opposite ends of the spectrum on the sentence involved in this particular case, they are crucially united in appealing to Turner's mental states at the time of the trial and sentencing to at least partly justify their competing positions.

Of course, there is no magic number of years after which an act is committed that we would say the mental states of a punishee fail to matter to the punishment deserved. Some trials take place decades after a crime is committed, and it is not unusual for judges, prosecutors, and defense attorneys to point to who the defendant currently is to defend a particular sentence. In 2014, for instance, a psychology professor, Norma Patricia Esparza, was offered a six-year prison sentence in exchange for pleading guilty to voluntary manslaughter for her purported role in the 1995 murder of a man who she

said raped her. This deal was made nearly 20 years after the crime was committed, and the district attorney's office made clear that Esparza's mental states were crucial to the offering of the plea, saying that it reflected Esparza's "acknowledgement and acceptance of her role in the victim's murder" ("Professor Admits Role in Killing of Her Alleged Rapist," 2014). Once again, acknowledgement and acceptance crucially involve mental states, such as beliefs, and the district attorney is talking about Esparza's states now, not 20 years earlier.

Perhaps the clearest and most powerful example in the criminal justice system on behalf of premise (2) is the evidence taken to be relevant at parole hearings or resentencing hearings. Some of the key conditions for release or reduced sentences include conceding guilt (Siegel and Ozug 2016),[9] good conduct in prison, completing classes, having been sufficiently rehabilitated, and not posing a danger to society. All of these involve mental states to varying degrees. It is, for instance, most natural to understand someone's being rehabilitated and not posing risks to society in terms of changes in their beliefs, desires, intentions, emotions, and dispositions. For instance, gang members once thought to be friends might now seem to be exploitative, bouts of anger might be channeled into education and advocacy rather than revenge, and the intention to be a role model for one's children might replace trying to impress one's peers.

We also see the relevance of current mental states in other punitive contexts. Compare two students known to have cheated: the first fully acknowledges that looking at her notes during an exam was wrong, is clearly contrite, and promises to never do so again, while the second flagrantly and steadfastly lies about it and shows no evidence that he won't cheat again. It is fairly standard for the second student's punishment to be harsher than that given to the first.

The third premise of the Transformation Argument is:

> (3) People can change, often in profoundly transformative ways, which can involve radical changes in their corresponding mental states.

Such transformations can be clearly seen by considering the two ends of the spectrum of life. On the early side, it is now widely known that the prefrontal cortex of the brains of adolescents and emerging adults is still developing, leading to their being more likely than adults to act on impulse, engage in dangerous or risky behavior, and misread social cues and emotions ("Teen Brain: Behavior, Problem Solving, and Decision Making," 2016). Indeed, "the frontal lobes, home to key components of the neural circuitry underlying 'executive functions' such as planning, working memory, and impulse control, are among the last areas of the brain to mature; they may not be fully developed until halfway through the third decade of life" (Johnson et al. 2009: 216). This fact by itself raises a host of questions about the level of responsibility that

[9] See also *New York Times* (n.d.).

adolescents and emerging adults bear for their actions, and the appropriate punishments that should be handed out to them. If, for instance, the underdeveloped brains of adolescents at least partly explain criminal behavior—behavior that wouldn't have occurred had they been adults—then holding them fully responsible for their actions, and punishing them as adults, seems wildly off the mark. But the point that I wish to emphasize here is that normal brain development often results in profound changes between adolescence and adulthood—changes that make the beliefs, decisions, and actions of the other seem foreign and perplexing. This is perhaps why adults often look back on some of the actions of their younger selves with embarrassment and even horror, and why teenagers frequently feel disconnected from, and poorly understood by, the adults around them.

On the later side, attention has been drawn to the fact that only 1% of serious crime is committed by people over the age of 60. According to Jonathan Turley, "Everyone agrees on what is the most reliable predictor of recidivism: age. As people get older, they statistically become less dangerous" (quoted in Tofig 1997). Turley refers to this period as "criminal menopause," a phenomenon where people lack the desire, and often the means, to engage in criminal activity. This raises serious questions about the rationale for punishment involving the elderly. If, for instance, there is abundant evidence that aging prisoners have been rehabilitated, and a negligible chance that they would pose a safety risk upon being released, then the skyrocketing numbers of elderly prisoners in the United States seem to cry out for explanation.[10] Again, though, the central point I wish to make here is the way in which this fact supports premise (3). For if even people who repeatedly engaged in criminal activity as young adults completely eschew this as they age, then there is a clear sense in which they have changed in transformative ways.

At the early end of the spectrum of life, then, there is the possibility that people might change; at the later end, there is the reality that they have changed. Both facts support premise (3).[11]

Moreover, there is a further kind of transformation that is importantly relevant here. In her recent book, L. A. Paul (2014) focuses on a phenomenon that she calls transformative experience, which involves experiences that are both epistemically and personally transformative. To be clear, I am distinguishing ordinary transformations, such as the typical changes that happen between adolescence and adulthood, from transformative experiences, which pick out the phenomena specifically described by

[10] Experts project that the elderly prison population in the U.S. will be over 400,000 in 2030, compared with 8,853 in 1981 ("At America's Expense: The Mass Incarceration of the Elderly," 2012).

[11] A recent article in *Quartz* has the headline: "You're a Completely Different Person at 14 and 77, the Longest-Running Personality Study Ever Has Found" (Goldhill 2017). The article discusses work recently published in *Psychology and Aging* that looks at six personality traits of subjects at 14 and then again at 77—self-confidence, perseverance, stability of moods, conscientiousness, originality, and desire to learn— and concludes that there is "no significant stability of any of the 6 characteristics" over the 63-year interval. This study makes particularly vivid the support for premise (3).

Paul. An epistemically transformative experience is one that provides new information that could not have been learned without having that kind of experience. A paradigmatic example here is that of Mary from Frank Jackson's (1982) well- known knowledge argument, who is a scientist specializing in the neurophysiology of vision. She acquires all of the physical information there is about color and related matters, but she herself has spent her entire life in a black-and-white room. Now imagine what happens when Mary leaves her room for the first time and sees a red apple. It seems plausible that, despite possessing knowledge of all of the physical information there is about color, she still learns something new—namely, what it is like to see color, and red in particular. Jackson uses this to conclude that physicalism is false, but what is relevant for our purposes is that Mary has an epistemically transformative experience. She learns what it is like to see red, which is something she couldn't have learned without an experience of this sort. While this is an extraordinary case, there are also plenty of ordinary examples of epistemically transformative experiences, such as what I would come to know were I to taste a durian fruit for the first time.[12]

An experience is personally transformative if it "changes you enough to substantially change your point of view, thus substantially revising your core preferences or revising how you experience being yourself" (Paul 2014: 16). In an important sense, personally transformative experiences change who you are by radically altering your point of view. "Such experiences may include experiencing a horrific physical attack, gaining a new sensory ability, having a traumatic accident, undergoing major surgery, winning an Olympic gold medal, participating in a revolution, having a religious conversion, having a child, experiencing the death of a parent, making a major scientific discovery, or experiencing the death of a child" (Paul 2014: 16). While personally transformative experiences are different than those that are merely epistemically transformative, this does not mean that there is not a central epistemic dimension to these. As Paul says, "if a personally transformative experience is a radically new experience for you, it means that important features of your future self, the self that results from the personal transformation, are epistemically inaccessible to your current, inexperienced self " (2014: 17).

Transformative experiences simpliciter, then, are those that are both epistemically and personally transformative in these senses. In addition to the examples Paul provides above, one experience that is frequently transformative for people and is particularly relevant for our purposes here is incarceration. For instance, in a very recent CNN article entitled "Ex-con Transforms to Entrepreneur Behind Bars," the author, Coss Marte, writes, "My personal transformation happened behind bars. I was sent to 'the box' after an altercation with a prison officer. After I was beaten, I was shoved into the cell and forced to do nothing but think. I asked myself 'Why?' How did I end up here?" (Marte 2016). Marte goes on to describe how his beliefs, desires, and actions radically changed after this moment. He began exercising, losing 70 pounds in

[12] Paul (2014) also discusses this example at length.

six months, he turned to God, he realized that selling drugs was wrong, and he "began to believe that [his] purpose was to give back instead of destroying individuals around [him]." This is not uncommon. Indeed, a recent article on "prison conversions" begins as follows: "The jail cell conversion from 'sinner to saint' or from nonbeliever to true believer is a well-known, indeed almost clichéd character arch [sic] in feel-good fiction, history, and media accounts" (Maruna et al. 2006: 161). The authors explain these conversions in terms of what psychologists William R. Miller and Janet C'deBaca (1994) call "quantum changes," which are sudden identity transformations that are "qualitatively different from the more common, incremental changes in human development" (Maruna etal. 2006: 161). Such quantum changes clearly involve radically new beliefs, such as those described by Marte above.

We are now in a position to see the Transformation Argument unfold as follows:

1. Punishment at a given time ought to be sensitive to the relevant evidence available at that time.

2. Punishment ought to be sensitive to the mental states of the one being punished, where this includes his or her mental states after the time of the punishable act.

3. People can change, often in profoundly transformative ways, which can involve radical changes in their corresponding mental states.

4. Strict, long-term punishments screen off any relevant future evidence, including the radically different mental states of people who have changed in transformative ways.

5. Therefore, strict, long-term punishments are epistemically irrational.

I have already offered defenses of (1)-(3), so let me say a few words about (4) and (5). Recall that strict punishments are those that are closed to revision, and thus it clearly follows that they screen off any non-exculpatory evidence beyond what is available at the time the punishment is handed out. Focusing on our paradigmatic case, natural life sentences say to all involved that there is no possible piece of information that could be learned between sentencing and death that could bear in any way on the punishment the convicted is said to deserve, short of what might ground an appeal. Nothing. So no matter how much a juvenile is transformed behind bars, and no matter how unrecognizable an elderly prisoner is from his earlier self, this is utterly irrelevant to whether they should be incarcerated. Our absence of knowledge about the future, our ignorance of what is to come, our lack of a crystal ball, is in no way a barrier to determining now what someone's life ought to be like decades from now.

While natural life sentences are the clearest example here, the same considerations hold for all strict punishments, such as any versions of truth-in-sentencing that require those convicted of crimes to serve 100% of their sentences.

However, screening off the possibility of even considering future, non-exculpatory evidence in relation to punishment is irrational. For as we have seen, future evidence, such as the radically different mental states of the punishee, can be relevant to the punishment that is deserved. This is especially clear when we consider all of the support on behalf of (3) showing that transformations while incarcerated are not only possible but likely. Indeed, consider this: if we take two defendants with different mental states regarding their crimes as deserving of different punishments at the time of sentencing, why would we not regard two stages of the same person—one at 19 and another at 49—with radically different attitudes toward his crime, as deserving of different punishments? Current selves and future selves can vary from one another no less than two altogether distinct people do.[13]

There is a related worry here that is worth mentioning briefly that appeals to the plausible moral principle, due to Aristotle, that *likes ought to be treated alike.*[14] If, for instance, two students produce work identical in quality, justice and fairness seem to require that I assess them comparably, just as I should sanction them similarly if these same two students are then found to have cheated under the same conditions. To fail to do so would be to violate this principle by not treating like cases alike—a violation at the heart of much discrimination that is deeply problematic. But now notice: there is no reason why moral principles of this sort should apply only once rather than in an ongoing way. Otherwise put, we ought to treat likes alike across time. Suppose, for instance, that A received a 40-year sentence while B received a 20-year sentence for the same crime. It may be perfectly appropriate to punish A more harshly than

[13] It should be clear that similar issues arise with respect to the death penalty. This point is made particularly vivid in a recent article in *The Atlantic,* "A Deadly Question: Have Juries Sentenced Hundreds of People to Death by Trying to Predict the Unpredictable?" The author, Abbie Vansickle, writes: "Before jurors could sentence someone to death [in Texas], they must first decide if the person will be a future danger. The precise wording of the question is convoluted, asking jurors 'whether there is a probability that the defendant would commit criminal acts of violence that would constitute a continuing threat to society.' At its core, it contains an incredible idea: Can we predict whether or not a killer will kill again?" (Vansickle 2016).

Vansickle goes on to show that experts are in agreement that the answer to this question is a negative one: "Dr. Mark Cunningham, a Seattle-based psychologist, and Dr. John Edens, a psychologist at Texas A&M University, have devoted their professional lives to the question of whether we can predict the future dangerousness of those convicted of crimes. Both have published extensively on the topic. And both have reached much the same conclusion. 'Juries show absolutely no predictive ability whatsoever,' Cunningham said. 'And, in fact, experts are similar.' The American Psychiatric Association ... has taken a similar position and implored the Supreme Court to ban the future dangerousness question in capital cases, saying in an amicus brief that '[t]he unreliability of psychiatric predictions of long-term future dangerousness is by now an established fact within the profession.' The APA concluded that the 'future dangerousness' question relies on junk science, and found that experts are wrong in two out of three predictions of 'future dangerousness'" (Vansickle 2016).

Given the possibility of transformations and transformative experiences, answers to the question about future dangerousness might be even more misguided, as past actions will not be a reliable guide to future ones when later selves are radically different from their earlier selves.

[14] See e.g. Winston (1974).

B at T1 because, though they committed the same crime, A's mental states crucially differ from B's at T1. Perhaps A takes delight in the memory of the crime or vows to commit more such crimes in the future, while B does not. At T2, however, suppose that A transforms or undergoes a transformative experience that renders his current mental states relevantly similar to the way B's were at T1. If A and B continue to be punished differently by serving sentences that vary significantly in length, then likes are not being treated alike.

But the problem that I want to raise here is an epistemic one: strict, long-term punishments screen off the possibility of treating likes alike. In particular, given premise (2) of the Transformation Argument, treating likes alike in relation to punishment requires sensitivity to the punishee's current mental states. However, strict, long-term punishments are incapable of taking this information into account. Thus, even if there are moral questions about the scope and detail of treating likes alike, this problem does not depend on settling them. For what is at issue here is an epistemic barrier to even aiming to treat likes alike, regardless of what this aiming fully amounts to. In other words, so long as we accept premise (2), strict long-term punishments close the door to treating likes alike under any interpretation, and hence close the door to satisfying a deep and powerful principle of morality.

4. The Transformative Choice Argument

As we saw, Paul develops the notion of a transformative experience, which involves an experience that is both epistemically and personally transformative. Such experiences, Paul argues further, raise problems for making rational choices involving them. This can be seen by considering a decision-theoretic framework. When calculating expected utilities within such a framework, two factors are taken into account: our subjective probabilities and our preferences. Suppose, for instance, that I am deciding whether to run a marathon. If I am calculating expected utilities, I should consider both the subjective probabilities that I would succeed, or fail, in running the marathon, and my preferences regarding doing so or not. I should decide to run the marathon, on this framework, if my calculations yield that this is the thing to do because it has the highest expected utility.

But Paul argues that there is a problem when decisions involve transformative experiences. Suppose, for instance, that I'm deciding whether to have a child for the first time. According to Paul:

> [such an example brings] out two problems with our ordinary way of making important personal decisions from our subjective perspective when the decisions involve transformative choices.
>
> First, transformative choices involve epistemically transformative experiences, compromising our ability to rationally assign subjective values to

> radically new outcomes. The subjective value of the lived experience that is the outcome of choosing to undergo the new experience is epistemically inaccessible to you, and this results in a type of ignorance that standard decision- theoretic models are ill-equipped to handle.
>
> Second, because of the personally transformative nature of the experience, your preferences concerning the acts that lead to the new outcomes can also change. In particular, having the new experience may change your post-experience preferences, or change how your postexperience self values outcomes. Transformative choices, then, ask you to make a decision where you must manage different selves at different times, with different sets of preferences. Which set of preferences should you be most concerned with? Your preferences now, or your preferences after the experience?" (Paul 2014: 47-8)

The Problem of Transformative Choice, then, involves two key dimensions: first, since there is epistemic impoverishment with respect to the nature of a transformative experience prior to having it, there is an ignorance of truths that compromises our ability to rationally assign subjective values to the outcomes in question. For instance, prior to tasting a durian fruit or having a child, I am ignorant of the nature of this gustatory experience and of what it is like to be a mother, and so my ability to calculate expected utilities within a decision-theoretic framework is severely limited. Second, transformative experiences can bring about a radical change in our preferences. These preferences might be unknown to me, which is another layer of ignorance that compromises our ability to calculate expected utilities. But even where I might know that my preferences will radically change, there is still the other issue that Paul raises: whose preferences do I take into account, mine now, or my future self's?[15]

I now want to consider the Problem of Transformative Choice in relation to the decision of whether to impose on someone a strict, long-term punishment, such as a natural life sentence. Suppose that I'm a judge and I'm deciding what the "appropriate punishment" is for a defendant by calculating the relevant expected utility of imposing a natural life sentence or not. Let's remain as neutral as possible on what precisely an "appropriate punishment" amounts to, but one feature that I defended in the previous section is that it has to be sensitive to the current mental states of the person being punished. Given that there is ample evidence that people often radically change while they're incarcerated for lengthy periods, either through typical transformations that occur between adolescence and being elderly or because of a transformative experience, there is a crucial ignorance of truths on my part as the judge: who will this person be in 10, 20, or 30 years? In particular, in calculating the expected utility of imposing a

[15] While Paul situates her argument within a decision theoretic framework, I do not wish to here take on board all aspects of such a framework. Instead, what is crucial for my purposes is the epistemic impoverishment involved when transformative experiences are concerned, which Paul's discussion very

natural life sentence, I need to consider the subjective probability that I will succeed, or fail, in appropriately punishing the defendant. But if the future mental states of the punishee are partially determinative of whether the punishment is appropriate, and relevant mental states can change dramatically over time via transformations and transformative experiences, then I am in a state of critical ignorance when making my decision. Hence, we have a parallel of the first dimension of the Problem of Transformative Choice.

It is interesting to emphasize the first-person and third-person aspects of this parallel. In Paul's original Problem of Transformative Choice, the decision is a first- person one: the fact that I might undergo a transformative experience compromises my ability to calculate the expected utility of my choice. Here, however, the decision is a third-person one: the fact that someone else might undergo a transformative experience compromises my ability to calculate the expected utility of my choice about that person's life. While both versions crucially involve epistemic impoverishment that compromises my ability to assign the relevant subjective values to the outcomes, the object of the ignorance varies: my own experiences versus another's.[16]

The other dimension of this problem involves the radical change in preferences that can come about via either transformations or transformative experiences, and we again find a parallel with decisions about imposing strict, long-term punishments. Preferences regarding appropriate punishments can, and do significantly change.[17] Some are the result of what I have been calling mere transformations, such as slow and steady changes in social attitudes, while others might be more properly regarded as transformative experiences, but both are relevant here. Let's begin with an example of the former: social attitudes regarding incarceration have radically changed in the past 20 years or so—a fact that might be powerfully illustrated by the transformation of the views of both Bill and Hillary Clinton on the matter. In her The New Jim Crow: Mass Incarceration in the Age of Colorblindness, Michelle Alexander writes:

> [I]n 1992, presidential candidate Bill Clinton vowed that he would never permit any Republican to be perceived as tougher on crime than he. True to his word, just weeks before the critical New Hampshire primary, Clinton chose to fly home to Arkansas to oversee the execution of Ricky Ray Rector, a mentally impaired black man who had so little conception of what was about to happen to him that he asked for the dessert from his last meal to be saved for him until the morning. After the execution, Clinton remarked, "I can be nicked a lot, but no one can say I'm soft on crime." (Alexander 2012: 56)

nicely develops.

[16] Of course, this version of the puzzle arises only when the possibility of the person in question changing radically is relevant to my decision, which is clearly the case when sentences are being imposed.

[17] In addition, values can change dramatically over time, at both individual and collective levels, and these can be crucial dimensions of appropriate punishments.

After being elected, Clinton signed into law the well-known (Violent Crime Control and Law Enforcement Act of 1994), which included, among various other things, $9.7 billion in funding for prisons, a significantly expanded federal death penalty, and mandated life sentences for criminals convicted of a violent felony after two or more prior convictions, including drug crimes (this was known as the "three-strikes" provision) (Violent Crime Control and Law Enforcement Act of 1994"). In 1996, Hillary Clinton gave a speech in New Hampshire in support of this Act, where she made her now infamous comment about "superpredators":

> [we] have to have an organized effort against gangs ... just as in a previous generation we had an organized effort against the mob. We need to take these people on. They are often connected to big drug cartels, they are not just gangs of kids anymore. They are often the kinds of kids that are called superpredators—no conscience, no empathy. We can talk about why they ended up that way, but first, we have to bring them to heel. (Graves 2016)

These sorts of attitudes led to sentences becoming harsher and longer, resulting in skyrocketing incarceration rates in the United States. The decisions were based in part on preferences regarding appropriate punishment by lawmakers, judges, prosecutors, and the broader society at large.

Recent years, however, have seen a dramatic shift in attitudes regarding criminal justice, with it now being widely agreed that earlier views about sentencing were simply wrong. Last year, Bill Clinton acknowledged that the 1994 crime bill was problematic in various ways: "I want to say a few words about [criminal justice reform]. Because I signed a bill that made the problem worse and I want to admit it" (Levitz 2015). And, of course, Clinton isn't alone; a recent headline on NPR reads, "20 Years Later, Parts of Major Crime Bill Viewed as Terrible Mistake" (Johnson 2014). Importantly, this acknowledgement is not simply the result of seeing that the crime bill contributed directly to the problem of mass incarceration; it is also due to a fundamental transformation of attitude: "Criminal justice policy was very much driven by public sentiment and a political instinct to appeal to the more negative punitive elements of public sentiment rather than to be driven by the facts" (Johnson 2014). Hillary Clinton, too, has now called for an end to mass incarceration, distanced herself from much of the 1994 crime bill, and has expressed regret about her "superpredators" comment, saying, "Looking back, I shouldn't have used those words, and I wouldn't use them today" (Gearen and Phillip 2016).

Judges and victims' families, too, experience transformations that lead to radical changes in attitudes and preferences. In a recent case, a judge testified at a hearing that he wrongly convicted a defendant of murder because he was "prejudiced during the trial":

> "As I read [a transcript of the original trial], I couldn't believe my eyes," the former judge said in an interview. "It was so obvious I had made a mistake. I got sick. Physically sick."
>
> Mr. Barbaro's change of heart led to a highly unusual spectacle this week in a Brooklyn courtroom: He took the witness stand in State Supreme Court to testify at a hearing that his own verdict should be set aside. (McKinley 2013)

In another case, Jeanne Bishop, the sister of a woman who was murdered, along with her husband and unborn child, wrote a book about forgiving the man who committed the murders, David Biro. Bishop writes:

> Many people—there is no shortage of them—are willing to write off the David Biros of the world. I was one of those people. Here was our argument: Look at what he did! It's so evil, so depraved, that only a malignant heart could have concocted it. He is without feeling. Any remorse he might express later is only a sham. He will never change.
>
> It is not true. I know this from my own transformation. God changed my heart. Why not the heart of David Biro? Why not the hearts of the thousands of people languishing in prison who have committed crimes for which we are willing to lock them up forever, without a second thought? (Bishop 2015: 152)

These passages represent radical changes in attitudes and preferences at some of the most crucial levels of the decision-making process about punishment within the criminal justice system—lawmakers, judges/juries, victims' families, and society more broadly. It is difficult to say precisely what caused these radical changes, or whether they involved gradual transformations, transformative experiences, or some combination thereof. But regardless of the causal origin, they raise a pressing, epistemic problem for making decisions about strict, long-term punishments, one that parallels the second dimension of Paul's Problem of Transformative Choice— namely, that we might be deeply epistemically impoverished with respect to our future preferences regarding appropriate punishment and, even if we're not, there is the crucial question: whose preferences do we take into account, ours now, or those of our future selves?

Notice, however, that the two dimensions of the Problem of Transformative Choice do not thereby lead to the conclusion that any decisions made in the face of the corresponding epistemic impoverishment are thereby epistemically irrational. It may not, for instance, be irrational to taste a durian fruit, even in a state of ignorance about both the "what it is like" of such an experience and one's own future preferences regarding it. But in the case of imposing strict, long-term punishments on people, the case is different. While I will not attempt to generate general epistemic principles

here for when, and only when, epistemic impoverishment leads to irrationality, I will highlight several features that distinguish tasting durian fruit from our topic at hand.

First, when handing out strict, long-term punishments, especially lengthy prison sentences, the stakes are high, and this can bear on whether the action in question is rational. There is no need to grant a thesis as controversial as pragmatic encroachment, according to which the standards for knowledge can vary, depending on the stakes, in order to maintain that stakes can bear on questions of epistemic rationality.[18] For instance, it may be irrational for me to not exercise greater epistemic caution around peanuts than you do when my child has a life-threatening allergy to them and yours does not. This doesn't necessarily mean that given the same evidence, you know, say, that there aren't peanuts in this slice of cake while I do not. Rather, it might be the case that we both know that there aren't peanuts in the cake but, given the incredibly high stakes, I need a grade of knowledge closer to certainty to act on this knowledge while you do not.[19] Similarly, there might be a greater degree of epistemic irrationality when the door is closed to relevant evidence in high-stakes situations than in low- stakes ones. Imposing a strict one-day punishment of no dessert, for instance, is far less irrational than a natural life sentence is, even though neither is open to revision, and at least one reason for this is that the stakes are much higher in the latter than they are in the former. The same is true in the case of durian fruit: the stakes are so low that it might be rational for me to taste it, even while I'm in a state of epistemic impoverishment about such an experience and my future preferences regarding it. But clearly the same isn't true when talking about depriving a fellow human being of freedom for decades.

Relatedly, the decision in question crucially involves the lives of others, not just of myself. This feature can be built into the high stakes nature of imposing strict, long-term punishments, but it is worth highlighting in its own right. It may, for instance, be adequate for me to quickly check that my life-jacket is on properly before sailing, but it might be necessary for me to double- or triple-check that this is so when I'm responsible for my child or yours.

Finally, there is a simple practical matter: some decisions are final, and so there is no possibility of leaving the door open to revision in light of new evidence. For instance, once one has a child, there is no turning back, even if one later discovers that it is not what one hoped it would be. In contrast, there is a very easy solution to the epistemic impoverishment we face when deliberating about punishments: don't make them strict. Recognize that profound and relevant changes can happen at every level of the process, and that our ignorance of the future should lead us to avoid binding our present selves when we don't have to. Indeed, even choices that we expect to significantly constrain our future selves, such as marriage, can be revisited on the basis of new evidence. This is precisely why divorce is legal.

[18] For a classic defense of pragmatic encroachment, see Stanley (2005) and Fantl and McGrath (2009).

[19] See Reed (2010) for this view.

We are now in a position to see that there is a slightly modified version of the Transformation Argument here, which we may call the Transformative Choice Argument. While transformations and transformative experiences are relevant in both arguments, this title emphasizes that the structure of this argument models Paul's involving transformative choice. Like the earlier argument, however, the Transformative Choice Argument brings to light the epistemic irrationality of strict, long-term punishments. There are two different and possibly even interacting ways in which transformations and transformative experiences enter the epistemic picture here. On the one hand, the fact that the punishee might radically change leads to an ignorance of truths on the part of the punisher, thereby compromising the punisher's ability to rationally assign subjective values to the outcomes in question. On the other hand, the fact that the punisher might radically change leads to an ignorance about which preferences ought to be taken into account when calculating expected utilities within a decision-theoretic framework. Moreover, we can certainly imagine these experiences interacting. Imagine, for instance, David Biro's experience upon hearing that the sister of his victims not only forgave him for the murders, but actually chose to cultivate a relationship with him. In such a case, Jeanne Bishop's transformative experience might be a catalyst for David Biro's own transformative experience. Given these two areas in which ignorance might compromise the rationality of decisionmaking, it is seriously epistemically problematic to make high-stakes decisions that are closed to the possibility of future revision. Strict, long-term punishments are, therefore, irrational.[20]

It is crucial to note that this argument depends on premises already defended in Section 3, such as (1) and (3). The main difference, and what has been the primary focus of the arguments in this section, is that premise (2) needs to be broadened to include not only the mental states of the punishee after the time of the punishable act but also those of the other relevant parties, such as the lawmakers, judges/juries, victims' families, and society more broadly. Rather than being understood as an entirely different argument, then, it is best to think of the Transformative Choice Argument as a modified or extended version of the Transformation Argument, according to which:

1. Punishment at a given time ought to be sensitive to the relevant evidence available at that time.

[20] It should also be noted that some decisions are made by groups or collectives, such as juries and states, and it is an open question how to understand transformations and transformative experiences within this framework. There are, for instance, metaphysical questions about what it means for groups to transform or have transformative experiences in the relevant sense. Does this mean merely that some individual members of the group have the experiences in question, or does something need to take place at the level of the group? There are also epistemic questions about how the possible changes in preferences of the member of a group bear on their decision-making. If, say, a small percentage of the group's members undergo a transformative experience that changes their preferences, does this then change the group's preferences? While these are fascinating questions, they lie beyond the scope of the present chapter.

2.) Punishment ought to be sensitive to the mental states of the punishees and punishers, where this includes their mental states after the time of the punishable act.[21]

3. People can change, often in profoundly transformative ways, which can involve radical changes in their corresponding mental states.

4. Strict, long-term punishments screen off any relevant future evidence, including the radically different mental states of people who have changed in transformative ways.

5. Therefore, strict, long-term punishments are epistemically irrational.

While only the second premise explicitly differs between the Transformation Argument and the Transformative Choice Argument, it should be clear that my development of Paul's framework also provides further support for (3) and (4). By focusing on the possibility of transformative experiences at many levels of the punitive process, and by providing examples of how the mental states of people at these levels bear on the appropriateness of the punishments in question, we have a deeper and more expansive understanding of how people can change over time, and how strict-long-term punishments screen off this relevant evidence in problematic ways.

5. Objections and Replies

I will now consider several objections to the arguments offered in this chapter, and provide some responses to them.

One theme that runs through many of the arguments is that we should not close epistemic doors to future evidence that might bear on the appropriateness of a given punishment, particularly the mental states of the punishee. This is perhaps most evident in my defense of premises (1) and (2) of the Transformation Argument. However, we have a number of practices, both legal and social, that close epistemic doors to varying degrees in ways that might seem to be at odds with these premises. Consider statutes of limitations: in many states, for instance, the statute of limitation on defamation is one year. Given this, even if new evidence is uncovered regarding a purportedly defamatory claim two years after it was first made public, it cannot be used for a lawsuit because of the statute of limitation. But then doesn't this seem to fly in the face of premises (1) and (2), according to which punishment at a given time ought to be sensitive to the evidence available at that time, including the mental states of the punishee? In particular, one who has defamed another cannot be punished for doing so in many states beyond a year, even if there is new evidence that bears on it. Doesn't

[21] I am understanding "punishers" here broadly to include lawmakers, judges/juries, victims' families, and society more broadly.

this, then, screen off the possibility of punishment tracking the available evidence in such cases?

The short answer to this question is yes, it does, but this need not undermine premises (1) and (2). The reason for this is that the justification for such statutes of limitation might be compatible with them being problematic in other ways. To see this, notice that statutes of limitation are longer the more serious the injury or crime, with some crimes, such as murder, having no statute of limitation at all. At least one of the explanations for this is that a lack of closure places a significant burden on those who might be open to claims or charges, and this needs to be weighed against the seriousness of the original matter. For instance, the stakes in at least many defamation cases are fairly low, with minimal damages, when compared with the burden of people living under a threat of a defamation lawsuit for the rest of their lives; but clearly such a burden is outweighed when talking about murder. In an effort to achieve the most just system overall, then, the law must function at a level of generality that isn't always optimal in other ways. Thus, there may be times when our practices close epistemic doors—in violation of (1) and (2)—because this is what is regarded as all-things-considered best. However, this doesn't mean that screening off available evidence isn't problematic in other ways, such as epistemically and morally.

A second objection that might be raised to the view defended here is that I seem to be endorsing a problematically weak conception of punishment. For instance, suppose that someone commits a particularly heinous crime and has a transformative experience within minutes of being arrested. Or suppose that there is a transformative experience pill that brings about a radical change in preferences upon taking it, and someone takes the pill immediately after committing an act of extreme violence. Am I saying that in both of these sorts of cases, the person in question shouldn't be punished because he or she had a radical change in relevant mental states?

By way of response to this objection, notice, first, that none of my arguments here depends on the strong thesis that only the mental states of the person being punished matter for whether a punishment is warranted. Rather, I am committed to the weaker thesis that mental states are a factor in determining appropriate punishment. Hence, even if a transformative experience pill, or less extraordinary means, brings about a radical shift in, say, the punishee's attitude toward his or her actions, there may still be other reasons why punishment is called for, such as retribution, justice for the victim or the victim's family, functioning as a deterrent to others, and so on. Moreover—and this is meant to be merely suggestive—the history of a transformation or a transformative experience might matter, and hence the fact that radical changes are directly brought about through manipulation might diminish the bearing it has on whether a given punishment is warranted. There are instructive parallels to draw on here, as it is often noted that historical factors matter to autonomy (Mele 1995), responsible agency (Fischer and Ravizza 1998), and moral accountability and moral character (Kapitan 2000). Suppose, for instance, that I don't want to do the hard work involved in cultivating virtues so I instead take a "courage pill." Even if this leads to my

having dispositions to behave in ways typical of those who are courageous, it may be doubted that I in fact have courage or that I have the right kind of courage to ground moral responsibility. This is because the origin of my character trait matters, either to whether I have it in the first place, or to the way in which it reflects normative facts about me. A similar line might be run here: a transformative experience pill might not be adequate for bearing on the punishment a punishee deserves because it might have the wrong kind of history. This might be made particularly vivid by imagining someone who is otherwise remorseless setting out to commit a cruelly violent act against another, counting on popping a transformative experience pill after the fact to reduce his or her prison sentence.[22]

A third objection is why I focus on strict, long-term punishments rather than merely strict ones. In particular, if the central objection to such punishments is that they screen off potentially valuable epistemic information, why wouldn't the problem arise with respect to all punishments that are not open to revision, regardless of whether they are long-term?

There is a sense in which this objection is correct that strictness is the feature of punishments that shoulders much of the epistemic significance here. In particular, the epistemic irrationality of the punishments at issue lies primarily in their screening off potentially relevant evidence by being closed to revision. But this doesn't mean that being long-term has no bearing on the rationality at all. First of all, the longer the punishment, the more time there is for transformations or transformative experiences to occur in all involved. Consider, for instance, the fact mentioned earlier that only 1% of serious crime is committed by people over the age of 60. Given this, a strict 40-year sentence on a 20-year-old is far more likely to screen off epistemically relevant evidence than is one imposed for two years on a 20-year-old. The long-term nature of the first punishment, then, makes it more epistemically objectionable than the second, even though they are both strict. Moreover, as noted earlier, stakes can bear on questions of epistemic rationality insofar as there might be a greater degree of epistemic irrationality when the door is closed to relevant evidence in high-stakes situations than in low-stakes ones, even without endorsing pragmatic encroachment. Again, as mentioned above,

[22] The example in the text involves popping a transformative experience pill after a punishable act, but similar questions might be asked about pre-emptive punishment. This calls to mind a scene from Philip Pullman's The Amber Spyglass:

> "I propose to send a man to find her and kill her before she can be tempted."
>
> "Father President," said Father Gomez at once, "I have done pre-emptive penance every day of my adult life. I have studied, I have trained—"
>
> The President held up his hand. Pre-emptive penance and absolution were doctrines researched and developed by the Consistorial Court, but not known to the wider church. They involved doing penance for a sin not yet committed, intense and fervent penance accompanied by scourging and flagellation, so as to build up, as it were, a store of credit. When the penance had reached the appropriate level for a particular sin, the penitent was granted absolution in advance, though he might never be called on to commit a sin. It was sometimes necessary to kill people, for example: and it was so much less troubling for the assassin if he could do so in a state of grace. (Pullman 2001: 75)

imposing a strict one-day punishment of no dessert is far less irrational than a natural life sentence is, even though neither is open to revision, and at least one explanation here is the difference in stakes.

6. Conclusions

Many types of arguments have been leveled against long-term punishments, especially natural life sentences. Economic ones focus on the ballooning costs of mass incarceration and the toll this takes on government budgets, especially as the age and medical expenses of prisoners rapidly increase. Legal ones ask whether such sentences are cruel and unusual and therefore violate the Eighth Amendment, particularly for juveniles. Social arguments ask whether natural life sentences discourage reform by providing no incentive for rehabilitation. Moral concerns are grounded in the dignity and rights of agents, while psychological objections call attention to the myriad causes of deviant behavior and their responsiveness to appropriate treatment. But what has been absent from these conversations is an epistemic argument that has to do with us—with our inability to say now what someone will be like years from now, especially when transformations and transformative experiences are a possibility.

In this chapter, I have attempted to fill this gap in the discussions. I have argued that strict, long-term punishments are epistemically irrational, with natural and de facto life sentences being the paradigmatic examples. Of course, this doesn't mean that, as a matter of fact, no one should ever serve lengthy prison sentences. Instead, the point is that no matter what the offense is, the door ought to be left open for revisiting the punishment in light of new evidence, particularly that brought about through radical changes in mental states.[23] In the criminal justice system, this means that it is irrational for the possibility of parole to be taken off the table at the outset of any sentence. This will promote not only practices that are epistemically proper, but also practices that treat those being punished with the respect that they deserve.[24]

References

Alexander, M. 2012. *The New Jim Crow: Mass Incarceration in the Age of Colorblindness.* New York: New Press.

"At America's Expense: The Mass Incarceration of the Elderly." 2012. <https:// www.aclu.org/ files/assets/elderlyprisonreport_2oi2o6i3_i .pdf>

Audi, R. 1997. "The Place of Testimony in the Fabric of Knowledge and Justification." *American Philosophical Quarterly* 34: 405-22.

[23] This chapter draws on some of the arguments I made in Lackey (2016).

[24] For very helpful comments on earlier versions of this paper, I am grateful to Nilanjan Das, Lauren

Audi, R. 1998. *Epistemology: A Contemporary Introduction to the Theory of Knowledge.* London: Routledge.

Bergmann, M. 1997. "Internalism, Externalism and the No-Defeater Condition." *Synthese* 110: 399-417.

Bishop, J. 2015. *Change of Heart: Justice, Mercy, and Making Peace with My Sister's Killer.* Louisville, KY: Westminster John Knox Press.

BonJour, L. 1980. "Externalist Theories of Epistemic Justification." *Midwest Studies in Philosophy* 5: 53-73.

Bonjour, L. 1985. *The Structure of Empirical Knowledge.* Cambridge, MA: Harvard University Press.

BonJour, L., and E. Sosa. 2003. *Epistemic Justification: Internalism vs. Externalism, Foundations vs. Virtues.* Oxford: Blackwell.

Burge, T. 1993. "Content Preservation." *Philosophical Review* 102: 457-88.

Burge, T. 1997. "Interlocution, Perception, and Memory." *Philosophical Studies* 86: 21-47.

Chisholm, R. M. 1989. *Theory of Knowledge.* Englewood Cliffs, NJ: Prentice Hall.

"DNA Exonerations in the United States." n.d. <https://www.innocenceproject.org/ dna- exonerations-in-the-united-states/>

Douven, I. 2006. "Assertion, Knowledge, and Rational Credibility." *Philosophical Review* 115: 449-85.

Fantl, J., and M. McGrath. 2009. *Knowledge in an Uncertain World.* Oxford: Oxford University Press.

"Fast Stats." 2016. National Center for Health Statistics, November 15. <https:// www.cdc.gov/ nchs/fastats/life-expectancy.htm>

Fischer, J., and M. Ravizza. 1998. *Responsibility and Control.* Cambridge: Cambridge University Press.

Fricker, E. 1987. "The Epistemology of Testimony." *Proceedings of the Aristotelian Society*, supp. vol. 61: 57-83.

Fricker, E. 1994. "Against Gullibility." In B. Krishna Matilal and A. Chakrabarti (eds), *Knowing from Words,* 125-61. Dordrecht: Kluwer Academic.

Gearen, A., and A. Phillip. 2016. "Clinton Regrets 1996 Remark on 'Super-Predators' After Encounter With Activist." *Washington Post* (February 25). <https:// www.washingtonpost. com/news/post-politics/wp/2016/02/25/clinton-heckled-by-black-lives-matter-activist/?utm_term=.25d2d976bca4>

Goldberg, S. 2017. "Should Have Known." *Synthese* 194: 2863-94.

Goldhill, O. 2017, February 19. "You're a Different Person at 14 and 77, the Longest-Running Personality Study Ever Has Found." <https://qz.com/914002/ youre-a-completely-different- person-at-14-and-77-the-longest-running-personality-study-ever-has-found/>

Goldman, A. I. 1986. *Epistemology and Cognition.* Cambridge, MA: Harvard University Press.

Graves, A. 2016. August 28. "Did Hillary Clinton Call African-American Youth 'Superpredators'?" https://www.politifact.com/truth-o-meter/statements/2016/aug/28/reince
Hawthorne, J. 2004. *Knowledge and Lotteries.* Oxford: Oxford University Press.
Hawthorne, J., and J. Stanley. 2008. "Knowledge and Action." *Journal of Philosophy* 105: 571-90.
Jackson, F. 1982. "Epiphenomenal Qualia." *Philosophical Quarterly* 32: 127-36.
Johnson, C. 2014, September 12. "20 Years Later, Parts Of Major Crime Bill Viewed As Terrible Mistake." <https://www.npr.org/2014/09/12/347736999/20-years-later-major-cr
Johnson, S. B., R. W. Blum, and J. N. Giedd, J. N. 2009. "Adolescent Maturity and the Brain: The Promise and Pitfalls of Neuroscience Research in Adolescent Health Policy." *Journal of Adolescent Health* 45: 216-21.
Kapitan, T. 2000. "Autonomy and Manipulated Freedom." *Philosophical Perspectives* 14: 81-103.
Kvanvig, J. L. 2009. "Assertion, Knowledge, and Lotteries." In P. Greenough and D. Pritchard (eds), *Williamson on Knowledge,* 140-60. Oxford: Oxford University Press.
Lackey, J. 2007. "Norms of Assertion." *Nous* 41: 594-626.
Lackey, J. 2008. *Learning from Words: Testimony as a Source of Knowledge.* Oxford: Oxford University Press.
Lackey, J. 2016. "The Irrationality of Natural Life Sentences." *New York Times* (February 1).
<https://opinionator.blogs.nytimes.com/2o16/o2/o1/the-irrationality-of-natural-life
Levin, S., and J. C. Wong. 2016. "Brock Turner's Statement Blames Sexual Assault on Stanford 'Party Culture.'" *The Guardian* (June 8). <https://www.theguardian.com/us-news/2016/jun/ 07/brock-turner-statement-stanford-rape-case-campus-culture>
Levitz, E. 2015. "Bill Clinton Admits His Crime Law Made Mass Incarceration 'Worse.'" MSNBC (July 15). <http://www.msnbc.com/msnbc/clinton-admits-his-crime-bill-made- mass-incarceration-worse>
Marte, C. 2016. "Ex-Con Transforms to Entrepreneur Behind Bars." CNN (October 4). <https://www.cnn.com/2016/10/04/us/iyw-coss-marte/>
Maruna, S., L. Wilson, and K. Curran. 2006. "Why God Is Often Found Behind Bars: Prison Conversions and the Crisis of Self-Narrative." *Research in Human Development* 3: 161-84.
McDowell, J. 1994. "Knowledge by Hearsay." In B. Krishna Matilal and A. Chakrabarti (eds), *Knowing from Words,* 195-224. Dordrecht: Kluwer Academic.
McKinley, J. C., Jr. 2013. "Ex-Brooklyn Judge Seeks Reversal of His Verdict in 1999 Murder Case." *New York Times* (December 12). <https://www.nytimes.com/2013/12/13/ nyregion/ex-brooklyn-judge-seeks-reversal-of-his-verdict-in-1999-murder-case.html>

Leydon-Hardy, Trevor Nyman, Laurie Paul, Kathryn Pogin, and Baron Reed.

McKinnon, R. 2013. "The Supportive Reasons Norm of Assertion." *American Philosophical Quarterly* 50: 121-35.

Mele, A. 1995. *Autonomous Agents.* Oxford: Oxford University Press.

"Michigan Life Expectancy Data For Youth Serving Natural Life Sentences." n.d. <https:// www.fairsentencingofyouth.org/media-resources/research-library/>

Miller, W. R., and J. C'deBaca. 1994. "Quantum Change: Toward a Psychology of Transformation." In T. F. Heatherton and J. L. Weinberger (eds), *Can Personality Change?,* 253-80. Washington, DC: American Psychological Association.

Neta, R. 2009. "Treating Something as a Reason for Action." *Nous* 43: 684-99.

New York Times, *The*. n.d. "N.Y./Region: The 'Innocent Prisoner's Dilemma.'" Video. <https://www.youtube.com/watch?v=JgbEnjnLtgc>

Nozick, R. 1981. *Philosophical Explanations.* Cambridge, MA: Belknap Press.

Olson, D. E., M. Seng, J. Boulger, and M. McClure. 2009. "The Impact of Illinois' Truth-inSentencing Law on Sentence Lengths, Time to Serve and Disciplinary Incidents of Convicted Murderers and Sex Offenders." Illinois Criminal Justice Information Authority. <http:// www.icjia.state.il.us/assets/pdf/ResearchReports/FINAL%20REPORT%20The%20Impact %20of%20Illinois%20Truth-in-Sentencing%20Law%200609.pdf>

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

Plantinga, A. 1993. *Warrant and Proper Function.* Oxford: Oxford University Press.

Pollock, J. 1986. *Contemporary Theories of Knowledge.* Totowa, NJ: Rowman and Littlefield. "Professor Admits Role in Killing of Her Alleged Rapist." 2014. CBS News (September 15).

<>https://www.cbsnews.com/news/professor-admits-role-in-killing-of-her-alleged-rap

Pullman, P. 2001. *The Amber Spyglass.* London: Scholastic Press.

Reed, B. 2006. "Epistemic Circularity Squared? Skepticism about Common Sense." *Philosophy and Phenomenological Research* 73: 186-97.

Reed, B. 2010. "A Defense of Stable Invariantism." *Nous* 44: 224-44.

Siegel, R., and M. Ozug. 2016. February 18. "From a Life Term to Life on the Outside: When Aging Felons Are Freed." NPR (February 18). <https://www.npr.org/2016/02/18/ 467057603/from-a-life-term-to-life-on-the-outside-when-aging-felons-are-freed>

Stanley, J. 2005. *Knowledge and Practical Interests.* Oxford: Oxford University Press.

Svrluga, S. 2016. "'I Take Him at His Word:' Judge in Stanford Sexual Assault Case Explained Controversial Sentence." *Washington Post* (June 14). <https://www.washingtonpost.com/ news/grade-point/wp/2016/06/14/i-take-him-at-his-word-judge-in-stanford-sexual-assault-case-explained-controversial-sentence/?noredirect=onandutm_term=.fe464d00c11a>

"Teen Brain: Behavior, Problem Solving, and Decision Making." 2016. AACAP (September). <https://www.aacap.org/AACAP/Families_and_Youth/Facts_for_Families/FFF-Guide/The- Teen-Brain-Behavior-Problem-Solving-and-Decision-Making-o95.aspx>

Tofig, D. 1997. "Aging Behind Bars." *Hartford Courant* (February 18). <https:// www.courant. com/news/connecticut/hc-xpm-1997-02-18-97O218o121-story.html>
Tuerkheimer, D. 2014. *Flawed Convictions: "Shaken Baby Syndrome" and the Inertia of Injustice.* Oxford: Oxford University Press.
Vansickle, A. 2016. "A Deadly Question: Have Juries Sentenced Hundreds of People to Death by Trying to Predict the Unpredictable?" *The Atlantic* (November 19). <https://www. theatlantic.com/politics/archive/2016/11/a-deadly-question/ 508232/>
Violent Crime Control and Law Enforcement Act. 1994, October 24. <https:// www.ncjrs.gov/ txtfiles/billfs.txt>
Williams, M. 1999. *Groundless Belief: An Essay on the Possibility of Epistemology.* Princeton, NJ: Princeton University Press.
Williamson, T. 2000. *Knowledge and its Limits.* Oxford: Oxford University Press.
Williamson, T. 2005. "Contextualism, Subject-Sensitive Invariantism and Knowledge of Knowledge." *Philosophical Quarterly* 55: 213-35.
Winston, K. I. 1974. "On Treating Like Cases Alike." *California Law Review* 62: 1-39.

14. Either/Or: Subjectivity, Objectivity and Value[(17)]

Katalin Balog

> If the doors of perception were cleansed every thing would appear to man as it is, Infinite. For man has closed himself up, till he sees all things thro' narrow chinks of his cavern.
>
> William Blake, *The Marriage of Heaven and Hell* (1794)

1. Introduction

My concern in this chapter is the role of subjectivity in our pursuit of the good. I propose that subjective thought as well as a subjective mental process underappreciated in philosophical psychology—contemplation[1]—are instrumental for discovering and apprehending a whole range of value. In fact, I will argue that our primary contact with these values is through experience, and that they could not be properly understood in any other way. This means that subjectivity has central importance in our evaluative life.

I understand subjectivity and objectivity in concepts, thoughts, and mental processes in terms of their connection to experience. A concept, thought or mental process is subjective to the degree it is connected to experience; whereas it is objective to the degree it is abstracted away from it.[2] If, as I argue, experience is required for the apprehension of a sort of value, it stands to reason that the subjective mental process I call "contemplation"—which involves voluntary attention to experience—together with subjective thought in the form of observational beliefs about value, are required for reaching a better, fuller, more vivid and accurate understanding of many of the values

[1] I use "contemplation" in a technical sense to be explained later in the chapter.

[2] This distinction goes back at least to the philosophies of Descartes and Kant. It has been recently discussed by Brian Loar (1987; 1995; 1997; 2003) and Thomas Nagel (1974; 1979; 1986), among many others.

[(17)] Katalin Balog, Either/Or: Subjectivity, Objectivity and Value In: Becoming Someone New: Essays on Transformative Experience, Choice, and Change. Edited by: Enoch Lambert and John Schwenkler, Oxford University Press (2020). © Katalin Balog.
DOI: 10.1093/oso/9780198823735.003.00015

that enter into our lives. This entails that cultivating subjectivity has great practical significance.

The plan is as follows. First, I provide a preliminary discussion of the subjective/objective distinction, especially as it applies to concepts and thought. Second, I outline the view that a vast array of values is apprehended in experience. Third, I provide an account of contemplation. In the final section, I elaborate on the role of subjectivity in pursuing the good.

2. Subjectivity and Objectivity in Concepts and Thought

I want to note at the outset that by "subjective" and "objective" I do not mean the familiar epistemological distinction that associates subjectivity with bias and motivated thought and objectivity with the standards of rationality. In an important article, Thomas Nagel (1979) spells out the difference between subjective and objective concepts in the following way. A concept is subjective if it is closely tied to perceptual or sensory experience. Examples are introspective concepts, such as the concept of feeling cold formed while undergoing the experience, perceptual demonstratives ("that beautiful thing"), or recognitional concepts ("rainbow," "blood," etc.). Subjective concepts can only be acquired and used by subjects that are familiar with the experiences involved, who know, in Nagel's expression, "what it's like" to have these experiences. A blind person, for example, cannot form a subjective concept of red as it presupposes a familiarity with how red appears.

On the other hand, even blind people can acquire the concept of red as a certain kind of light reflectance. This concept is objective. Objectivity is marked by abstraction from sensory and perceptual sources of information. Concepts like "citizen" or "collateral damage" are objective; as are concepts such as "charge," "electron," "logarithm," etc. which are the farthest removed from experiential concepts. Intelligent creatures with a very different sensorium might not share any of our subjective concepts while they could well entertain our more abstract ones. "Objective" and "subjective" does not denote a binary difference between concepts; rather, it denotes a continuum that runs from more subjective to more objective concepts, with perceptual demonstratives providing the subjective pole, and the concepts in mathematics and fundamental physics providing the objective pole.

Based on this distinction, we can draw a further distinction between subjective and objective thought: A thought is subjective or objective corresponding to the degree that it employs subjective or objective concepts. For example, experiencing beauty might prompt us to make the subjective observational judgment "That is beautiful" in the same way experiencing red prompts the subjective observational judgment "That

is red." On the other hand, thinking of red in scientific terms is an example of objective thought.

Recently Laurie Paul, in her book Transformative Experience, has argued persuasively that engaging in subjective thinking is indispensable for evaluating possible scenarios, and that because of this, we need to substantially rethink our understanding of decision-making. By placing my ideas in the context of Paul's work on decision-making, I will be better able to explain the scope of my project. Paul introduces the concept of transformative decision—a decision at least one of whose outcomes involves a "transformative experience." A transformative experience is both epistemically transformative—in that it involves experience hitherto unknown to the subject, which, as a result of undergoing the experience, will become familiar to them—and personally transformative—in that it substantially changes the subject's values and aspirations. The more radically life-altering the decision is—think of, for example, having a child—the harder to imagine what life will be like afterwards, the harder it is to have a subjective conception of what life with a child involves, the more likely it is that the experience is both epistemically and personally transformative. Ahead of the decision, Paul says, one has no way of forming a subjective concept of the emotional bond that will develop, or of the intense focus on the welfare of one's child that will replace one's former free-wheeling, self-centered existence. Moreover, Paul thinks (2015: 12) that experience has subjective value, grounded in the phenomenal character of the experience, which needs to be taken into account in decisions and which, similar to the phenomenal character of the experience, can only be properly grasped—for the purposes of authentic decision-making—in a subjective way. This has particular consequences for transformative decisions.

In the case of transformative decisions, she argues, one cannot be both rational and authentic. Authenticity requires that one is able to grasp the relevant subjective values from one's own point of view—that is, in subjective terms that alone can fully reveal their significance for the subject. A rational assessment of the values of outcomes in a transformative decision would have to include in its purview the subjective values of the experiences involved. But even if one were able to receive objective information about those subjective values (say, from statistical evidence) and make a rational decision, that decision would not be authentic in the absence of an appropriate subjective grasp of the values at stake.

I share Paul's concern with the role of subjective thought in grasping values. But my focus is not on decision-making itself but on the role subjective thought plays in the process of coming to know what one values—a precondition of making rational decisions, and in general, of identifying and moving towards meaningful projects and relationships. As Talbot Brewer (2009) observes:

> [Some] conceptions [of decision-making] encourage the thought that one's outlook on value ought rationally to be complete and determinate before one begins the work of deliberation about particular cases. They provide

> no insight into the most fundamental work of deliberation, which is the formation and revision of a tenable conception of the good. Instead, they reduce deliberative rationality to skill in estimating probabilities and performing calculations. But the work of proper deliberation is not the work of an accountant. It is the work of a seeker after the good, and it often requires a fresh straining to form a more just and palpable sense of the goods bearing on one's ongoing activities. (p. 103)

Paul thinks about experience as itself having what she calls subjective value. Eating kiwi is a good thing, at least partly because the experience of eating kiwi has value for me. I suggest that the main contribution of the experiential life to the evaluative life is not that experience itself is valuable, but that it represents, in addition to perceptual properties, value properties as well. We come to appreciate the value of things, activities, and people, their appeal, their depth and significance, etc.—wholly or partly—in our everyday phenomenal experience.[3,4] This is not to deny that experience has value in itself—pleasure and pain are obvious examples. But—*pace* hedonistic utilitarianism[5]—pleasure is not the sole, or even the main, kind of good in the world. It is the flower that is beautiful, not my experience of it, though the experience has some value in itself as well.

3. Value and Experience

> Contemplated by different people, [the] same glass can be a thousand things, however, because each man charges what he is looking at with emotion, and nobody sees it as it is but how his desires and state of mind ... see it. (Bunuel 1958)
>
> All *primordial* comportment toward the world ... is a *primordial emotional* comportment of value-ception. (Scheler 1973 [1916]: 229)

I will now make some fairly general remarks about value and experience. I am going to be only as specific as is necessary to establish the significance of subjective thought in our evaluative lives. I propose that in perception, our experience represents evaluative properties as well as strictly perceptual ones. There is a distinction between the perceptual/sensory aspects of experiential content and its affective aspects. More is presented in, say, a visual experience than the standardly agreed upon perceptual properties, such as color, shape, illumination, motion, and their co-instantiation in objects.[6] The flower appears to me pale blue, fragrant, with a sharply defined shape; it

[3] Johnston (2001) has a similar view. See also Noordhof (2018).

[4] By "phenomenal experience" here I mean sensory and perceptual, as well as emotional states. Though I think intentional states such as beliefs have their own distinctive phenomenology, they don't count as experiences in the way I intend to use the term here.

[5] For a powerful argument against hedonism, see Nozick (1974).

[6] For arguments that perceptual experience represents "higher-level" properties, such as natural kinds, or causal relations, see Siegel (2006), Block (2014).

also appears delicate, refreshing and delightful. Such sensuous features give the world its significance at the most fundamental level. Freshness or beauty is as much part of the content of my experience of the flower as is its color and shape. In experience, we grasp the fully determinate versions of determinables such the beautiful, the kitschy, the sublime and the horrific, the appealing and repulsive. Intuitively, such values can only be grasped by people who have experienced instances of them. Mark Johnston[7] calls these the "inherently sensuous" values. Some philosophers hold that in addition to the inherently sensuous ones, broadly moral values, such as kindness, probity, or shiftiness, are also represented in perception.[8]

Some philosophers, on the other hand, hold that values are represented not in sensory/perceptual experience, but in the emotions that attend them. This is most plausible for the moral realm, with e.g. anger, resentment, or love representing their objects as offensive, despicable, or desirable.[9] But one might argue that even sensuous values, such as beauty or ugliness, are represented in the attending "aesthetic emotions" and not in the perceptual experience itself.

It is hard to adjudicate this question, as perception and affect are closely intertwined in experience. Though I prefer the view on which perceptual states possess an affective component—rather than merely occurring concomitant with emotions—I will be noncommittal about whether values are represented in perception, emotion, or both. My only claim is that there is, as Scheler calls it, "value-ception," i.e. that value is represented experientially, in the broad sense that includes emotions and sentiment. The evaluative aspect of experience is often quite salient. But even when that is not the case, even when it is not obvious or easy to discern, it is still present. Perhaps all normal experience is evaluative—things don't tend to be experienced as entirely neutral.[10]

Are all values apt to be represented in experience? For example, can values such as the revelatory power of conceptual art, or the wrongness of perjury, be experienced? Take the case of witnessing perjury while—through simultaneous fact-checking— one is also aware that this is indeed perjury. There might be—if one is well-brought- up—an immediate, felt sense of wrongness to one's apprehension of the situation. But is this "felt sense" purely an experience? On the one hand, it might be that a conceptual grasp of the fact that one is witnessing perjury penetrates one's (perceptual or emotional) experience,[11] and produces a phenomenal representation of wrongness.[12] But it might

[7] Johnston (2001: 18 2).

[8] McDowell (19 8 5), Johnston (2001), Audi (2018), Noordhof (2018).

[9] See e.g. Doring (2003).

[10] I also wish to remain neutral about a host of meta-ethical questions about the nature of value. Any meta-ethical position that is compatible with the claim that value is represented in experience—and, as will argue later, that it could not be represented in thought in the absence of evaluative experience—will do.

[11] For an account of cognitive penetration in the moral case, see Cowan (2015).

[12] One of the main arguments for the perceptual representation of higher-level properties is the Contrast Argument (see Siegel 2006). For an application in the aesthetic case, see Stokes (2018).

also be that the experience by itself does not represent the wrongness of perjury, that additional conceptual thought is required for its grasp.[13] I will not take up this issue. Either way, I maintain that a broad range of values are represented in experience.[14]

But is experience really required for grasping the sensuous values? Could it not be adequately grasped in objective thought alone? To the degree that this was the case, subjective thought about values might be dispensable. To bring the question into focus, consider the following thought experiment. In a twist on Frank Jackson's famous Knowledge Argument/[15] imagine that there is a person—we might call her Insensate Mary—whose experience, though it has the same sensory/perceptual content as ours, lacks an affective aspect altogether. When she sees a roadside accident, she has no "gut reaction" to it: she feels no aversion, no horror, no sadness, nor (as the case might be) morbid curiosity.

The thought-experiment relies on the idea that we can conceive of an experience that has perceptual content but lacks affective content altogether. It seems likely that these aspects can vary somewhat independently. We have all experienced walking the same streets or looking at the same objects experiencing them wildly differently depending on our mood, or general state of mind. What was drab and uninspiring one day might be exciting another day. To take an example that is more directly relevant, consider the way morphine affects pain: It apparently it leaves the sensory content intact but removes the affective component, the awfulness of pain. Whether or not that is indeed the case, however, and so whether or not the two aspects of experiential content can really come apart, it seems possible to conceive of a situation in which they do.

Unlike Jackson's Mary, Insensate Mary does not have superhuman knowledge in objective terms—to ponder the abilities of such a creature would be irrelevant for my inquiry, which centers on the lives of humans. Insensate Mary is quite normal in her abilities to reason in objective terms but—unlike Jackson's Mary—has never experienced the myriad determinate ways in which something can be beautiful, or scary, or desirable. She cannot know about these things from the first-person perspective, she cannot think about them subjectively.

Could it turn out that these values are not "inherently" sensuous—in other words, that she could still form some other, objective conception of them, adequate for the purposes of practical engagement? This doesn't seem to be possible. Take for example an artist, Insensate Pollock. Just as his thinking about color needs to be based on his experience of color—a blind painter can't have grasp of color adequate for their

[13] E.g. Lyons (2018) thinks that this felt sense is actually something post-perceptual, a perceptual-seeming state which has experiential phenomenal character and conceptual content.

[14] There is good reason to think that the range of values that can be represented in experience is fairly wide. It is fairly plausible to assume that the many species of mammals that can discern love, harm, danger, untrustworthiness or unfairness—and which are capable of empathy, reconciliation or disapproval (see de Waal 2006)—do so through conscious perception or emotion rather than conceptual thought.

[15] Jackson (1982).

practical engagement of it on canvas—he cannot be insensate either. Never having experienced the "push and pull" of colors, shapes, and textures, it seems he cannot produce art—except by chance.

Here is why. Suppose that Mary could somehow form, on the basis of third-person information, the concept beauty* referring to beautiful things. It seems clear that it would not be a genuine value concept, fit for practical activity. Her evaluations of beauty would lack authenticity, the special first-person grasp we have of the sensuous values. Authenticity is required not only for a proper appreciation of value, but also for a proper engagement with it. If she were to seek out beautiful paintings, it would not be driven by her response to the intrinsic value of beauty, and so, in a certain way, it would be inexplicable.[16] Sensuous values can be authentically, properly grasped only through subjective experience—through the visceral experiences of attraction and recoil they produce in us. Their authority—their power to motivate—lies in the affect they produce in us. Affective insensitivity cannot be compensated by objectivity.

As a matter of fact, even in the case of normal subjects, forming an objective conception of some manifestation of sensuous value is, at least in some respects, defective for the purposes of practical life. Take a case where you form a second-hand conception of some manifestation of beauty. Suppose you had a twin whose tastes and evaluative dispositions with regard to the sensuous values are identical to yours. Even so, it would not do you much good to defer to her to make beauty judgments, say, about a work of art you haven't seen, or a new garden in the neighborhood that you don't want to bother to visit. Those judgments wouldn't be authentically yours, as they wouldn't reflect *your* actual take on the world. Moreover, since such deferential judgments would not be based in a sensuous grasp of an instance of a fully determinate version of the beautiful, you could only make a fairly general appraisal of the attractions of the garden or the art. Finally, such appraisals only have an indirect claim on you compared to the instances of value you encounter yourself.

How about the more abstract, objective value concepts such as justice or fairness? I suggest that Insensate Mary—lacking a sensuous appreciation of the world—lacks the necessary basis to form *any* evaluative concepts at all. I think that nothing could have significance or meaning for a creature who lacked basic experiences of sensuous value. I will discuss this in a little more detail later.

4. Contemplation and Subjective Thinking the way ... is to become subjective, ... to become subject.

(Kierkegaard 1992 [1846]: 109)

In the following pages, I develop an account of subjective thought to provide a psychological framework in which to understand what Brewer calls the "formation

[16] Paul (2015) makes this point in connection with the importance of authenticity in transformative

and revision of a tenable conception of the good." In many of his religious writings,[17] Kierkegaard discusses, either implicitly or explicitly, the difference between a subjective and objective thinker.[18] He uses the term "subjective" and "objective" to demarcate more than just a difference in two styles of conceptual thought. He seems to imply that mere conceptual thought—even subjective conceptual thought—doesn't make a thinker subjective. A subjective thinker cultivates subjective conceptual thought rooted in a mental process that I, for lack of a better word, will call *contemplation.*

Contemplation, in the sense I intend, is more than just having experience; not all experiencing counts as contemplation. It involves a distinct process that is, like conceptual thought, partly voluntary: the deployment of *attention* to the content of one's experience, for example, to the colors and textures of the ocean while swimming, or to the haunting melody of a song. Experiences being non-conceptual representations,[19] contemplation does not involve reasoning but associations among memories, images, fantasies and thoughts. In contemplation, experience is "held" in attention and is explored without a particular goal in mind. This is different from other forms of attention deployed in thought or perception in which it is fastmoving and task-oriented.

Contemplation happens in small ways every time we stop to appreciate the world as we experience it, every time we are present for what is happening in a deliberate fashion, rather than breezing through on automatic pilot (or being absorbed in thought to the exclusion of experience). Examples include appreciation of nature, reflection on art, or paying attention to other people in an experiential sort of way (as opposed to just thinking about their words). Contemplation is separate from conceptual thought, but it provides the basis for subjective conceptual thought, and in particular, the basis for conceptual thought about value. At the same time, conceptual distinctions made on the basis of experience serve to enhance and enrich our experience, as in the case of wine-tasting. Though values—like taste—are often perceived in experience, they are not always readily so—their discernment requires patient contemplation and subjective conceptual thought. Contemplation, contrary to a widely held misconception, does not necessarily mean an inward focus on experience itself, a drawing back of one's attention to the self. It can just as well consist in paying attention to the objects of experience, through the experience, rather than through conceptual thought.

---

decisions.

[17] See esp. Kierkegaard (1992 [1846]).

[18] Kierkegaard never explicitly defines these terms. I introduce what I think is a plausible elaboration of what Kierkegaard had in mind.

[19] Gareth Evans (1982) introduced the term "conceptual" vs. "non-conceptual" content into contemporary philosophical discussion where the latter characterizes perceptual and certain other mental states. His idea was that these states are like thoughts insofar as they are representational. But they are unlike thoughts in that where a thought, say *that the cat is hungry*, has a *conceptualized content* composed out of the concepts *cat* and *is hungry,* a person's visual perception of the cat is not composed out of concepts, and represents the cat in some other way.

Contemplation can happen in formalized contexts such as meditation[20,21] and certain kinds of psychotherapy as well.[22] Part of the aim of meditation and psychotherapy is to impart skills of contemplation that can be exercised in daily life in one's ongoing perceptual engagement with the world.

In extreme cases, there might be temporary states of pure contemplation, or states of pure conceptual thought,[23] but normally the two intermingle and interact. A thinker is subjective, in the way I propose to use the term, to the degree that they engage in contemplation and their conceptual thought tends toward the subjective side. A thinker is objective to the degree that they refrain from contemplation and their conceptual thought tends toward the objective side.

People differ in their style of thought. Some engage the world mostly by thinking conceptually, and tend to pay little attention to their experience. On the other hand, a highly contemplative person—for example, an artist or a monk—is attuned to their experience, and their conceptual thought is mostly subjective. Most people are somewhere in between those two extremes. In the novel The Brothers Karamazov, two of the brothers, Ivan and Alyosha, share the Karamazov "sensuality," a certain intensity of experience; but Ivan reacts by repressing his feelings and withdrawing into the stance of a cynical, witty observer, while Alyosha cultivates his inner life through religious training, and views the abstractions of politicians and "learned people" as "wicked nonsense."

Thought is understood primarily as conceptual. I want to stress that there is another mode of thought—experiential and non-conceptual—that plays an

important role in our psychology. While its best-known practices—meditation and psychotherapy—are generally understood as tools for stress management or relief from psychological pain for those in need, I suggest that contemplation has a broader role to play. In the next section, I elaborate on some of the ways contemplation is essential in our evaluative life.

[20] The two main kinds of meditation involves focused attention, usually on the breath or parts of the body, and open monitoring, which involves an evenly hovering attention over the field of conscious experience. See Lutz et al. (2008) for the distinction.

[21] See e.g. Lutz et al. (2015), and Grossenbacher and Quaglia (2017) for an account in cognitive neuroscience of various contemplative practices.

[22] Various forms of psychoanalysis encourage this kind of thinking, e.g., in "free association" (Freudian analysis) or "active imagination" (Jungian analysis).

[23] A scenario like that is described in Dennett (1978) involving complete sensory deprivation.

5. Reasons to Cultivate Subjectivity in the Pursuit of the Good

5.1. Cultivating Subjectivity to Have a Better Grasp of Values

In Sinan Antoon's[24] novel on the Iraq war, *The Corpse Washer,* the protagonist describes his job in this way:

> If death is a postman, then I receive his letters every day. I am the one who opens carefully the bloodied and torn envelopes. I am the one who washes them, who removes the stamps of death and dries and perfumes them mumbling what I don't entirely believe in. Then I wrap them carefully in white so they may reach their final reader—the grave.

The corpse washer, as a function of his occupation, attains a more direct, subjective perspective on war's destruction than those who learn about it from the news. He sees and *contemplates*, over and over, the bodies, maimed, drained of life; he touches them, his seeing and sensing intertwined with his terror. His contemplation reveals not just the gunshot wounds, corpses, and destruction, but also their particular awfulness. Such perception of dreadfulness forms the basis of further associations—it brings up related memories and images of friends and family killed, maimed, and exiled. Attention to all this throws the dreadfulness of war and violence into sharp relief, so it can be reflected and acted upon.

Sensibility can be trained in many ways—art, music, mountain-climbing, even corpse-washing are ways to increase sensibility. It proceeds by joining contemplation with conceptual elaboration: The experience of an arpeggio is different for one who has the concept and for one who doesn't; war's ravages are different for someone who has the experience and vocabulary to grasp its gruesome details than for someone who doesn't. The more attention one pays, and the greater conceptual sophistication one has, the more fine-grained one's understanding of the world through experience, and the more fine-grained one's discernment of the values manifesting in it.

The corpse washer's experiences lead him to grasp the war's significance in ways that separates him from people who hear about it from the news. This kind of difference in understanding is manifest, for example, in debates between survivors of mass shootings and politicians in charge of gun laws. The corpse washer or the survivor thinks about war and violence subjectively, with a thorough understanding of the stakes involved, while most politicians, despite the occasional dramatic footage, think about it in objective,[25] impoverished, sanitized terms such as "civilian casualty" or "col-

[24] Antoon (2014).

[25] I want to remind the reader that by "objective" thought I don't mean unbiased or rational

lateral damage." Both ways of thinking have their virtues and vices. Politicians are expected to weigh competing considerations in an impartial manner, which requires a certain level of abstraction. Survivors, however, understand the stakes involved more thoroughly than people who never experienced violence.

5.2. Overcoming Abstraction, Bias, and Self-Deception

Contemplation, however, is difficult. It is often difficult to become aware of the affective valence of experience, as there are many other things that claim one's attention. Moreover, contemplation requires a certain amount of self-denial: It requires the acceptance of one's experience of the world as it is, with all the pain as well as joy it contains, instead of seeing it in the light most pleasant and flattering for the self. This runs against strong forces in human nature. Most commonly, one turns one's back on subjectivity to escape from pain. As Freud has described, the mind has powerful built-in mechanisms that helps expunge unwanted experience from consciousness: Repression, dissociation, sublimation, etc. In more ordinary cases, one simply takes a step back to consider the "facts" rather than dwell on the experience.

Even small discomfort can prompt one to turn away from experience. If I pass a homeless person on the street, I might tune out so as not to feel any pangs of guilt about not having contributed. We react strongly to anything that challenges our own self-image. What needs to be learned, as Iris Murdoch has observed,[26] is "how real things can be looked at and loved without being seized and used, without being appropriated into the greedy organism of the self." Contemplation requires the skill to direct attention to what's going on in an impartial way, without distortion—which is perhaps most difficult when it comes to understanding and empathizing with others.

There are many ways that different cultures developed to foster this kind of selfdiscovery. The Chinese book of divination, the *I Ching*, guides decisions not by providing practical solutions, but by offering an opportunity to contemplate one's experience, in memory and imagination, in an open-ended manner. Buddhist meditation practice is based on the understanding that contemplative engagement requires considerable effort. It provides training in sustaining awareness of experience without turning away from any aspect of it or trying to change it to something else. This has to be achieved in the face of the fact that conceptual thought is more likely to command one's attention, and often in its less fruitful displays, like rumination, day-dreaming, and the like, is driven by various agendas. Both traditions hold that to fully realize the meaning of one's life and to be able to best align with the good in it, one needs to live a contemplative life.

It might be objected that one can live a full life dedicated to values and ideals without dwelling on subjectivity at all. Value concepts differ in their subjectivity;

thought—I simply mean a kind of thought that tends toward abstraction.

[26] Murdoch (1970: 23).

some are more abstract than others. "Beauty" or "love" is subjective;[27]7 "vulgarity" is more objective; "injustice" is more objective still. One can perhaps ignore sensuous values and still have a rich evaluative life, have enough aspirations to occupy one for a lifetime, dedicated to promoting justice, equality, freedom.

Here is why objective thought, when it is not accompanied by a subjective engagement with value, is not enough. Any capacity to see the world through an evaluative lens—as I suggested earlier regarding the case of Insensate Mary—is grounded in experience. To the degree that one's evaluative life centers around abstract concepts, while having dimmed or deadened any perception of value, one has cut the vital source of one's connection to value. Life becomes cerebral and dull, one's self-knowledge lacking. Thinking about matters of value in abstract terms conceals the significance of the issues involved, as in the case of talk about "collateral damage." Just as Insensate Mary's concept beauty* is defective for the purposes of practical life, abstract conceptions of value such as "social justice," when not connected to vivid experiential conceptions of, say, pain, injury, disrespect, etc., are defective as well. They do not reveal why one should care at all.[28]

Moreover, basing one's aspirations on objective conceptions of value alone can prepare the ground for self-delusion and compulsive action. The dictum "Know thyself" doesn't just call you to find out about your capacities and dispositions, habits of thought, etc., though that is certainly part of it. It also calls for a more thorough familiarity with how you experience the world and yourself—with your perceptions of the world, sensations of your body, your emotions, memories, and fantasies as well. Only in this way can you find out what you really want—which of the goods that you perceive in the world really matter for you. For example, you might not realize until you start paying more attention that you like to be with a person very much even though the person's virtues are not what you "officially" care about most. Figuring this out requires conceptual thought, but it would not be possible without close attention to your experience. Objective conceptual thought cannot entirely replace contemplation and subjective thought in a pursuit of the good.

What about misrepresentation of value in experience? Does careful contemplation guarantee freedom from self-deception? Perhaps there are some sensuous values, such as beauty or ugliness, attractiveness or repulsiveness, whose exemplification is wholly dependent on the experiences of the subject. In that case, there could be no gap

[27] Love has both subjective and more complex, more objective conceptions; I will skip over that here.

[28] Indeed, contemplation may be a last resort for moral agency when normal moral thought had lost its purchase, as in situations of extreme dehumanization. In the film *Son of Saul* (Laszlo Nemes, 2015), the protagonist, Saul, a prisoner in Auschwitz, is first shown in a deadened state of being. In the opening scenes he hardly pays attention to any of the carnage happening around him—he is just doing his tasks in a mechanical way. He regains a sense of agency amidst the general chaos after witnessing, with his attention focused on it like a laser beam, the murder of one particular person, whom he later declares to be his son (the movie is ambiguous about whether that is so). He now has a mission: to try to arrange a proper funeral.

between what the subject experiences as beautiful and what is really beautiful for the subject. One would still need to be attentive so as not to make a mistake by inattention if one aspires to orient towards the sensuous goods; however, once proper attention is deployed, there would be no possibility of correction based on further evidence. But this is not true generally; there are kinds of value that are sometimes misrepresented in experience. I can, for example, be mistaken in my experience that someone is shifty, or honest, or vulgar. This means that the contemplation of experience is just the first step in a process of discovery, a "straining" to come to an accurate grasp of the values involved. A plain example is the way emotions and memories can distort experience; jealousy, for example, can prevent one from seeing the virtues of another person accurately. Contemplating not just the experience of this person as, say, being vulgar, but also the inner landscape that shades one's perception of this person in unflattering ways, paves the way to form better and more just conceptions of their behavior—of seeing them as they are.[29]

5.3. Cultivating Subjectivity and Appreciating the World

According to the Buddhist and Daoist traditions, contemplation fosters appreciation. The point is not just that contemplation discloses goods to pursue—it also leads to valuing life as such. It can bestow vividness and meaning on ordinary, boring, everyday activities; only it can create the sense that one's life has touched the world. As D. T. Suzuki (1956) remarks:

> Life, as far as it is lived in concreto, is above concepts as well as images. To understand it we have to dive into and come in touch with it personally; to pick up or cut out a piece of it for inspection murders it; when you think you have got into the essence of it; it is no more, for it has ceased to live but lies immobile and all dried up. (Suzuki 1956: 58)

The following poem by Czeslaw Milosz expresses the power of contemplation to create appreciation for life. The poem itself needs to be contemplated rather than just thought about to yield its meaning:

> At the entrance, my bare feet on the dirt floor, Here, gusts of heat; at my back, white clouds, I stare and stare. It seems I was called for this: To glorify things just because they are.[30]

[29] Murdoch (1970: 32) discusses a case like this involving a mother and a daughter-in-law. Throughout the book, she emphasizes the moral relevance of the inner life in coming to a more accurate perception of value. If successfully executed, "selfish concerns vanish, nothing exists except the things which are seen."

[30] From "Blacksmith Shop," by Czeslaw Milosz, in *Provinces* (New York: Ecco Press, 1991).

To unlock the poem, it is necessary to contemplate the feelings, memories, and images evoked by it. Such contemplation captures the deep structure of experience; uncovering its resonances and dissonances, its associative structure. Only then—and only partially—can one's understanding be put to words and thought about conceptually. The poem's imagery mixes with experiences you remember, creating the sense of awe it expresses, a sense of the value of ordinary life, discovered through a deep awareness of experience.

5.4. Cultivating Subjectivity and "Slow Decisions"

I began to have an idea of my life, not as the slow shaping of achievement to fitmy preconceived purposes, but as the gradual discovery and growth of a purpose which I did not know. (Milner 2011 [1934])

> The task of attention goes on all the time and at apparently empty and everyday moments we are "looking", making those little peering efforts of imagination which have such important cumulative results. (Murdoch 1970: 22)
>
> ... moral change and moral achievement are slow ... the exercise of our freedom ... is a small piecemeal business which goes on all the time and ... not something that is switched off in between the occurrence of explicit moral choices. (Murdoch 1970: 33)

Decisions can be framed more objectively, or more subjectively. One might decide between career choices by weighing them in relatively abstract terms, such as the pay involved, the security the job offers, or opportunities for learning. Alternatively, one might frame the decision, at least in part, in terms of an assessment of what it might be like to work in those occupations. Such a framing allows a more nuanced sense of the values involved as well as an authentic appreciation of them.[31]

Objective deliberation, with its straightforward considerations of value and chance of success, is relatively fast. It doesn't take weeks or months to complete. On the other hand, when value judgments are made in isolation from relevant experience, they often have limited power to change behavior, as the fate of many a New Year's resolution shows. Subjective deliberation, by contrast, takes time. In a slow, subjective decision process, one allows oneself to live with the question for a while, to dwell in experience long enough for one's feelings about the decision to emerge. How one wants to spend one's life and with whom one wants to spend it are discoveries rather than simply a matter of rationally appraising the known parameters of a situation.

Sometime, even when all the relevant facts are known, it takes time to make a decision that feels appropriate. Such a decision can happen as a result of a familiar

[31] Paul (2015) argues that only such decisions count as authentic.

consideration presenting itself over and over again—for example, in the case of having to deal with an untrustworthy friend or lover. You might have understood, in an abstract sense, what is happening, and may have been at the point of trying to draw the consequences. But your decision never stuck, it never felt quite right. However, once you have gradually allowed yourself to fully experience your friend's behavior—putting aside any effort to find excuses—you will likely come to a point where you can act. As Kierkegaard puts it in Either/Or:

> Ask yourself, and continue to ask until you find the answer. For one may have known a thing many times and acknowledged it ... and yet it is only by the deep inward movements, only by the indescribable emotions of the heart, that for the first time you are convinced that what you have known belongs to you ... for only the truth that edifies is truth for you.

The crucial part of decision-making is the process of discovery that precedes it. Whereas rational choice theory provides an algorithm for comparing choices once one knows their value, it is contemplation that helps determining what those values are, and making them vivid to pull one with enough motivating force. As Dostoyevsky says, in Notes from the Underground, "reason is a good thing, that can't be disputed, but reason is only reason and satisfies only man's intellectual faculties, while volition is a manifestation of the whole of life."

Paul casts doubt on the relevance of decision theory as a useful model for transformative decisions. My approach signifies another way in which taking subjective experience seriously suggests a shortcoming of rational decision theory as a model for actual decision-making. Decision theory treats knowledge of values as readily available, or at least has nothing to say about how to obtain it. In fact, finding out what one values is the most difficult, and most crucial, part of decision-making.

6. Conclusion

I have argued that value discourse benefits from incorporating the notions of contemplation and subjective thought in philosophical psychology. The upshot is that cultivating subjectivity is an important and underappreciated aspect of the pursuit of the good. According to Kierkegaard, it is the most important of all.

References

Antoon, Sinan. 2014. *The Corpse Washer*. New Haven, CT: Yale University Press.

Audi, Robert. 2018. "Moral Perception Defended." In A. Bergqvist and R. Cowan (eds), *Evaluative Perception*. Oxford: Oxford University Press.

Barnard, P. J., D. J. Duke, R. W. Byrne, and I. Davidson. 2007. "Differentiation in Cognitive and Emotional Meanings: An Evolutionary Analysis." *Cognition and Emotion* 21(6): 1155-83.

Barnard, P. J., and J. Teasdale. 1991. "Interacting Cognitive Subsystems: A Systemic Approach to Cognitive-Affective Interaction and Change." *Cognition and Emotion* 5(5): 1-39.

Block, Ned. 2014. "Seeing-As in the Light of Vision Science." *Philosophy and Phenomenological Research* 89(3): 560-72. DOI: 10.nn/phpr.12135

Brewer, Talbot. 2009. *The Retrieval of Ethics*. Oxford: Oxford University Press.

Bunuel, Luis. 1958. "The Cinema, Instrument of Poetry." *Cuadermos de la Universidad de Mexico* 4 (December).

Cowan, R. 2015. "Cognitive Penetrability and Ethical Perception." *Review of Philosophy and Psychology* 6(4): 665-82.

Dennett, Daniel. 1978. "Where Am I?" In *Brainstorms.* Cambridge, MA: MIT Press.

Doring, S. A. 2003. "Explaining Action by Emotion." *Philosophical Quarterly* 53(211): 214-30.

Evans, G. 1982. *The Varieties of Reference.* Oxford: Oxford University Press.

Jackson, F. 1982. "Epiphenomenal Qualia." *Philosophical Quarterly* 32: 127-36.

Johnston, M. 2001. "The Authority of Affect." *Philosophy and Phenomenological Research* 68: 181-214.

Grossenbacher, P. G., and J. T. Quaglia. 2017. "Contemplative Cognition: A More Integrative

Framework for Advancing Mindfulness and Meditation Research." *Mindfulness* 8(6): 1580-93. doi: 10.1007/s12671-017-0730-1

Kierkegaard, S0ren. 1992 [1846]. *Concluding Unscientific Postscript to Philosophical Fragments.*

Princeton, NJ: Princeton University Press.

Loar, Brian. 1987. "Subjective Intentionality." *Philosophical Topics* 1: 89-124. Repr. in *Consciousness and Meaning: Selected Essays by Brian Loar*, ed. Katalin Balog and Stephanie Beardman (Oxford: Oxford University Press, 2017).

Loar, Brian. 1995. "Reference from the First-Person Perspective." *Philosophical Issues: Content* 6: 53-72. Repr. in *Consciousness and Meaning: Selected Essays by Brian Loar,* ed. Katalin Balog and Stephanie Beardman (Oxford: Oxford University Press, 2017).

Loar, Brian. 1997. "Phenomenal States." In Ned Block, Owen Flanagan, and Guven Guzeldere (eds), The Nature of Consciousness. Cambridge, MA: MIT Press. Repr. in Consciousness and Meaning: Selected Essays by Brian Loar, ed. Katalin Balog and Stephanie Beardman (Oxford: Oxford University Press, 2017).

Loar, Brian. 2003. "Transparent Experience and the Availability of Qualia." In Q. Smith and A. Jokic (eds), Consciousness: New Philosophical Perspectives. Oxford: Oxford University Press. Repr. in Consciousness and Meaning: Selected Essays by Brian

Loar, ed. Katalin Balog and Stephanie Beardman (Oxford: Oxford University Press, 2017).
Lutz, A., A. Jha, J. Dunne, and C. Saron. 2015. "Investigating the Phenomenological Matrix of Mindfulness-Related Practices from a Neurocognitive Perspective." American Psychologist 70(7): 632-58. http://dx.doi.org/10.1037/a0039585
Lutz, A., H. A. Slagter, J. D. Dunne, and R. J. Davidson. 2008. "Attention Regulation and Monitoring in Meditation." *Trends in Cognitive Sciences* 12(4): 163-9. doi: 10.1016/j. tics.2008.01.005
Lyons, Jack. 2018. "Perception and Intuition of Evaluative Properties." In A. Bergqvist and R. Cowan (eds), *Evaluative Perception*. Oxford: Oxford University Press.
McDowell, J. 1985. "Values and Secondary Qualities." In T. Honderich (ed.), *Morality and Objectivity*. London: Routledge and Kegan Paul.
McDowell, J. 1994. *Mind and World.* Cambridge, Mass.: Harvard University Press.
Milner, Marion. 2011 [1934]. *A Life of One's Own.* Abingdon: Routledge.
Murdoch, Iris. 1970. *The Sovereignty of Good.* London: Routledge and Kegan Paul.
Nagel, Thomas. 1974. "What Is It Like to Be a Bat?" *Philosophical Review* 83: 435-50.
Nagel, Thomas. 1979. "Subjective and Objective." In *Mortal Questions.* Cambridge: Cambridge University Press.
Nagel, Thomas. 1986. *The View from Nowhere*. Oxford: Oxford University Press.
Noordhof, Paul. 2018. "Evaluative Perception as Response Dependent Representation." In A. Bergqvist and R. Cowan (eds), *Evaluative Perception*. Oxford: Oxford University Press.
Nozick, Robert. 1974. *Anarchy, State, and Utopia*. New York: Basic Books.
Paul, L. 2015. *Transformative Experience.* Oxford: Oxford University Press.
Scheler, Max. 1973 [1916]. *Formalism in Ethics and Non-Formal Ethics of Values: A New Attempt Toward the Foundation of an Ethical Personalism*. Evanston, IL: Northwestern University Press.
Searle, John. 1992. *The Rediscovery of the Mind.* Cambridge, MA: MIT Press.
Siegel, S. 2006. "Which Properties Are Represented in Perception?" In T. Gendler Szabo and J. Hawthorne (eds), *Perceptual Experience*. Oxford: Oxford: Oxford University Press.
Stokes, Dustin. 2018. "Rich Perceptual Content and Aesthetic Properties." In A. Bergqvist and R. Cowan (eds), *Evaluative Perception*. Oxford: Oxford University Press.
Suzuki, D. T. 1956. *Zen Buddhism.* New York: Doubleday Anchor.
Waal, Frans de. 2006. *Primates and Philosophers: How Morality Evolved*. Princeton, NJ: Princeton University Press.

15. Death: The Ultimate Transformative Experience[(18)]

Evan Thompson

1. Introduction

Death is the ultimate transformative experience. I do not mean the state of being dead, in which the person has ceased to exist and we suppose there to be no possibility of experience for that person. I mean the whole process of dying, culminating in the end of a person's life (Morison 1971: 694-8). So understood, death is "epistemically transformative," because you cannot know what it is like to die until you experience dying and this experience can enable you to understand things in a new way.[1] Death is also "personally transformative," because it changes how you experience yourself in ways that you cannot fully grasp before these changes happen. At the same time, death is unlike any other transformative experience. It is the ultimate one, not only in being final, inevitable, and all-encompassing, but also in having fundamental significance. It is the kind of transformative experience against which or from which all other transformative experiences can be viewed. Death's power to reveal new truths about your self and your life is exceptional. Dying comprises prospective and retrospective perspectives that differ from those of any other experience. A consideration of death as the ultimate transformative experience brings an important perspective to the philosophy of death while offering insights for physicians, nurses, hospice workers, and family members who care for dying loved ones.

My inspiration comes from a statement L. A. Paul makes in a footnote in Transformative Experience: "Your own death is the ultimate transformative experience, and as such, you are particularly ill-equipped to approach it rationally" (2014: 111, n. 6). Paul does not elaborate on this remark, but I propose to take two thoughts from what she writes in the context of the argument of her book.

[1] The terms "epistemically transformative" and "personally transformative" come from Paul (2014).

[(18)] Evan Thompson, Death: The Ultimate Transformative Experience In: Becoming Someone New: Essays on Transformative Experience, Choice, and Change. Edited by: Enoch Lambert and John Schwenkler, Oxford University Press (2020).

© Evan Thompson. DOI: 10.1093/oso/9780198823735.003.00016

First, death shares the characteristics that make transformative experiences as a whole pose a challenge to a standard (but culturally specific[2]) way of thinking about what it is to be rational and authentic. A transformative experience teaches you something you could not have learned without having that kind of experience, and it changes your desires, preferences, and understanding of your self. In Paul's view, the challenge to rational authenticity arises in situations of transformative choice, in which you must decide whether to undergo a transformative experience. (Two of her examples are whether to become a parent and whether to receive a cochlear implant if you have been deaf from birth.) Authenticity requires that you choose based on your own ability to imagine what the new experience will be like for you, and on the value you place on having that kind of experience. Rationality requires that you forecast the likely outcomes of your choice, including what the resulting new experiences will be like for you, that you assess the outcomes according to your own first-personal and "subjective values," and that you match your decision to your preferences (to what you prefer to happen as a result of your choice, given your subjective values).[3] The problem is that it seems that you cannot adequately imagine what the experience will be like before you undergo it, and that you cannot determine what value the experience will have for you, because the experience may change your values and hence your preferences. Death—the whole transformative experience of dying—seems to pose the same kind of problem: How are you supposed to be able to imagine it before you undergo it, and how are you supposed to know how it may change your values and your preferences, or—more generally—your sense of what is personally meaningful?

Nevertheless—and this is the second thought I take from Paul's remark—death is not just another kind of transformative experience. It is, as she says, the ultimate one. Although you can choose to end your own life, most of us do not choose our own deaths and you cannot choose not to die.[4] In the case of every other transformative experience, you focus on what you will be like after the experience, whereas in the case of death, the problem is how to approach it, given our limited knowledge of what it is like. Its inevitability and finality, and the dissolution it entails, are especially hard to grasp from within.[5] The barrier that death throws up to the imagination is of an entirely different order of difficulty from that of any other transformative experience. The affective response and the immediate engagement of the emotions that the real or imagined prospect of death elicits are unlike those elicited by the prospect of any other transformative experience. For these reasons, if you are generally ill-equipped

[2] Namely, reflecting the norms of WEIRD (Western, Educated, Industrialized, Rich, and Democratic) societies. See Henrich, Heine, and Norenzayan (2010).

[3] For important elaboration and clarification of the concept of subjective value, see Paul (2015a: esp. 477-9 and 500-03.

[4] For a powerful and moving statement of one person's reflective choice to take her own life, see Bennett (2014).

[5] For philosophical discussions of the apparent impossibility of grasping death from within, see Nagel (1986: 223-31) and Valberg (2007: pt 2, 'Death').

to approach transformative experiences rationally, you are particularly ill-equipped to approach death rationally.

These thoughts set the context for this chapter. I wish to examine the unique characteristics of the ultimate transformative experience of death. My main motivation is to bring philosophy to bear on the experience of death in hospice (including home hospice care). My conviction is that the current philosophical concern with transformative experience, which Paul's book has sparked, can help to reawaken a guiding impulse of ancient philosophy—namely, to prepare one's mind for death, whenever it may come, and to live one's life accordingly.[6] The hospice is a crucible for facing death. I wish to bring the hospice movement and philosophy together to learn from and invigorate each other. This aim sets the context and scope of much of my discussion. It is the reason that I focus mainly on the experience of dying from a terminal illness or from senescence in the modern Western hospice setting.

Nevertheless, I do not mean to imply that dying in hospice can serve as a model for understanding every kind of death. Indeed, one might wonder whether there is any such thing as *"the* experience of death." One might even ask: To what extent can we talk about "the experience of death" in general terms as an identifiable category of human experience about which we can make useful generalizations?[7]

From the perspective of the hospice movement, it has been very important to talk about death in this way in order to develop better models of how to care for dying people. From a broader perspective, however, I would submit that there may be no such thing as "*the* experience of death" in the sense in which there may be no such thing as *"the* experience of love" or *"the* experience of grief." There are many ways to experience death or love or grief. Nevertheless, in each case there remains an identifiable category of human experience about which we can make useful generalizations, as long as we recognize that there are many subcategories and that our generalizations must take into account the social, cultural, and historical settings of the experience, including how those settings can be constitutive of the experience. We also need to recognize what Ian Hacking (1996) calls "looping effects," whereby the categories that we use to classify people and their experiences change people and their experiences, especially in medical contexts. We must also not lose sight of the ineliminable individual variability of experience. More to the point philosophically, our conception of experience must be a "thick" or "wide" conception that includes the way the world is as part of the experience.

This chapter is part of a larger project that includes reflecting on what it might mean to be better or worse equipped to approach death rationally and authentically. That project includes an examination of whether you can imagine what it is like to die. I believe that you may be able to do so in certain respects, and that this fact has important implications for our understanding of transformative experiences. You need

[6] See Hadot (1995).

[7] I am grateful to Troy Jollimore for raising these concerns.

to rely on a range of first- and second-person information, including second- person testimony from the dying (including "moral testimony" about the values of experiences (Harman 2015)), your own experiences of caring for the dying, contemplative practices, and "thick" imagination—that is, imagination that engages your emotions, and targets not just the qualities of experience from within but also the way the world is beyond your individual self.[8] Literature is an especially important source for this kind of rich imagination. Consider Tolstoy's *The Death of Ivan Ilyich* or Herman Broch's *The Death of Virgil* or the Japanese tradition of *jisei* (the death poem). I also believe that being well-equipped to approach death rationally and authentically requires caring for the dying, and that our society is particularly ill- equipped to approach death rationally and authentically because caring for the dying is not a part of normal life.[9] As physician Haider Warraich writes in his recent and important book, *Modern Death,* "At no time in our history has death been farther from home than in the last few decades" (2017: 40).

Finally, I believe that whether an experience is transformative can depend causally and constitutively on matters of social justice (Barnes 2015). Social conditions can make the transformative experience of death more or less harmful. These social conditions include not only whether one has access to proper end-of-life care but also the pervasive medicalization of death. As Warraich notes, "contrary to general perception, never has death been as feared as it is today. The more medicalized death gets, the longer people are debilitated before the end, the more cloistered those who die become, the more terrifying death gets" (2017: 9). Thus, issues about social justice are inseparable from issues about how we are to be able to approach death authentically and rationally. This chapter, which focuses mainly on certain unique characteristics of the ultimate transformative experience of death, is a prelude to that larger project.

2. Defining Death

Many events make up the dying process. Which one counts as death depends on how we conceptualize or define death, and on the medical standards or criteria we use to determine when death has occurred.[10]

One event is the point at which dying is irreversible and death is assured. Stephen Luper (2016) calls this event "threshold death." It has been conceptualized in different ways, either as the irreversible cessation of the integrated functioning of the organism ("integration death": Luper 2016), as the irreversible loss of consciousness, or as the irreversible loss of personhood. Different medical standards or criteria have been proposed in order to determine when death occurs, according to one or another of these

[8] See Campbell (2015) and Paul (2015b: 806-13). See also Paul (2015a: 478).

[9] See Doughty (2015) and Nutik Zitter (2017).

[10] These definitions and medical standards and criteria are also inseparable from political and practical considerations about organ donation and transplantation. See Nair-Collins (2015).

conceptions. The main standards that have been debated are the irreversible cessation of circulatory-respiratory function, the irreversible cessation of the functioning of the entire brain (including the brain stem), and the irreversible cessation of the functioning of the cerebral cortex (DeGrazia 2017).

Another event is the ending of the dying process, which Luper (2016) calls *"denouement* death." Neurologist Steven Laureys describes that event as "the discontinuous event ... that separates the continuous process of dying from the subsequent disintegration" (2005: 900). It defines death in the restrictive sense of life's termination point.

These conceptualizations and definitions of death, as well as the medical standards for determining when death occurs, are based on analyzing death from the outside, in third-person terms, not from the inside, in first-person terms. By themselves, they tell us little about the experience of death.

Nevertheless, threshold death and dénouement death, as well as many other phases of dying, can be subjectively experienced, especially in the contemporary hospice setting and as a result of modern methods for resuscitation (Kellehear 2014). Medical sociologist Allan Kellehear writes in his book, The Inner Life of the Dying Person, that "studies suggest and describe transformative psychological, social, and spiritual experiences that occur, for some people at least, even into the first few minutes after a medical judgment of death has been pronounced" (2014: 195).

Two points deserve emphasis here. First, notice that Kellehear is talking about medical judgements, which are fallible. Judging that a person is dead is one thing; the person's actually being dead is another thing. Second, the transition from being alive to being dead can be indeterminate. Consider these words from a nurse's blog: "The first thing I learned was that alive/dead is not the easy dichotomy that you would think it is. Either you're dead or not. But I go uncommonly religious and tell people that the Bible says it's 'the hour of our death' for a reason. In the ER, I saw one person during my entire time there be pronounced and then sit up and pull her tube out. Now I know to give it a while before pronouncing" (Not Nurse Ratched 2016).

The grey zone between life and death is not necessarily a sign of a lack of scientific knowledge; on the contrary, it reflects modern medical advances and is characteristic of "modern death." In Warraich's words: "Not only have biomedical advances changed the ecology, epidemiology, and economics of death, but the very ethos of death—in the most abstract possible sense—has changed. Far from being clearer, the line between life and death has become far more blurry. These days we can't even be sure if someone is alive or dead without getting a battery of tests" (2017: 9).

Nevertheless, we do lack knowledge about one crucial matter—namely, about exactly how the brain supports consciousness and whether it is possible for certain kinds of conscious awareness to remain present for some time after the heart stops beating and breathing ceases. For example, a study in rats showed that complex dynamical patterns of brain activity can continue for up to 30 seconds after cardiac arrest (Borjigin etal. 2013), and a recent study in humans found that brain activity apparently

continued in one patient up to 10 minutes after the heart stopped beating (Norton etal. 2017). Consider also the state of very deep coma, which is considered to be the turning point between a living brain and a dead brain. The isoelectric (flatline) EEG that characterizes this state is one of the criteria used to assess brain death. A recent study found that brain activity is generated in the hippocampal formation (a structure crucial for memory and spatial cognition) and transmitted to the cortex during very deep coma in humans and cats, and that this brain phenomenon is deeper than the one reflected by the isoelectric EEG (Kroeger et al. 2013).

The passage from Kellehear (2014) I quoted earlier occurs in the context of his considering the related cases of patients who have been revived from coma or resuscitated after cardiac arrest and who report conscious experiences of thought and feeling that seem to them to have occurred while they were in these states. Kellehear describes these cases of "near-death experiences" as ones in which the patients "are actually aware of their own death" (2014: 195). But caution is needed here. First, the patients did not die (and accordingly we call their experiences "neardeath experiences," not "death experiences"). Second, it is not clear that these experiences happen at the precise time that the EEG is isoelectric, rather than just before the flatline state or just after this state, when the patients are recovering.[11]

In addition, we need to distinguish between being dead as a reversible medical condition and being dead as an absolute irreversible condition. Death in the medically reversible sense is equivalent to the cessation of blood flow, respiration, brain-stem activity, and whole-brain function. As a result of advances in resuscitation science, it has become possible "to reverse death [in this sense], not only in the immediate postmortem period but also potentially for relatively prolonged periods of time after it has occurred" (Parnia 2014: 76). The reason for this possibility is that "human cells do not become irreversibly damaged or die immediately postmortem" (p. 76). Thus, death as an absolute irreversible condition occurs only when cell death becomes permanent.

These considerations about the fallibility of medical judgments of death and the grey zone between being alive and being dead concern the difficulty of saying exactly when the end of dying happens and the state of being dead ensues. There is a corresponding difficulty about saying exactly when dying begins. Saying that dying begins when death is inevitable has obvious problems. On the one hand, death is inevitable once life begins. From this perspective, the process of living is the same as the process of dying. On the other hand, in certain cases of critical injury or illness, imminent death may seem inevitable, but the patient recovers; or imminent death may not seem inevitable (the patient has a good chance of recovery), but nonetheless the patient dies.

The hospice setting is relevant here. Hospice workers often use the terms the "preactive phase of dying" and the "active phase of dying."[12] The pre-active phase usually lasts on average a few weeks, and the active phase lasts on average a few days. These phases

[11] For further discussion, see Thompson (2015: 302-3).

[12] See e.g. Hospice Patients Alliance (n.d.).

have characteristic physical and psychological signs and symptoms, though there is a large amount of individual variability.

Given all these considerations, I am defining "death" as the whole process of dying, especially "active dying," including—but not limited to—its end point and the ensuing state of being dead.

I mean for this definition to be neutral on the question of the survival of consciousness after death. In particular, I do not wish to build into the definition of death an entailment to the metaphysical thesis of annihilationism. In other words, I would like the definition to be consistent with the possibility of some kind of continuance of the mind or consciousness after death, despite whatever doubts I may have about such continuance actually being the case. Whether experience in some sense is possible after absolute irreversible death is just the issue of the survival of consciousness after death, which I do not think the definition of death should prejudge. Whether experience in some form is possible during or after reversible medical death is the question raised by near-death experiences. I take this question to be open, given our limited knowledge of the brain and how it supports consciousness. Nevertheless, I think that the evidence to date from the study of near-death experiences does not give us reason to think that consciousness continues after death (Thompson 2015: ch. 9).

In summary, the point of defining "death" as the whole process of dying is to accommodate the experiential side of active dying and its culmination in the state of being dead. In contrast, philosophical discussions that follow the Epicurean definition of death as the state of dissolution or annihilation (e.g. Warren 2004), or that define death simply as the absence of consciousness and life (e.g. Edwards 1969), are too restrictive. They elide the dying experience and fail to include the dying person's inner life. We can remedy this shortcoming by bringing the concept of transformative experience to bear on our thinking about death. By describing death as the ultimate transformative experience, we can reinstate the first-person perspective in the philosophy of death.

3. The Transformative Experience of Dying

Death is epistemically transformative because it teaches you things you cannot learn until you undergo it, and it is personally transformative because it deeply changes how you experience your self. These facts are well documented by those who care for the dying and listen to what they have to say.[13]

A recent study in the Netherlands examined how cancer patients react to the realization that their death may be imminent (Yang etal. 2010). The realization provokes great distress, leading to what the authors of the study call an "existential crisis" with any or all of the following seven characteristics: (1) an acute awareness of one's own

[13] I draw here mainly from Kellehear (2014). See also Dowling Singh (1998) and Kuhl (2002).

finitude; (2) a limited sense of the future, including a feeling that what remains of the future is threatening and alarming; (3) a loss of meaning, especially for the sense of purpose in one's life; (4) fear, anxiety, panic, and despair; (5) extreme loneliness, even when surrounded by love and care; (6) powerlessness; and (7) an identity crisis, aggravated especially by physical mutilation and dependence on others.

Philosopher Ken Chung, a month before he died at the age of 39, published a blog essay, "Is Dying a Transformative Experience?," in which he reflects on Paul's conception of transformative experience in the light of his own experience of having Stage IV pancreatic cancer.

> So what tells me I'm dying? There's no gut feeling—no gut knowledge that I only have so much time. There's just my doctors' words buttressed by data that only they can intelligibly interpret ... But there is something to knowing that you have a disease that's going to kill you soon. If you accept that fact, and acknowledge it deep down and in all the things you do, I think it does transform you ... [T]here is this gap between me, who is dying, and you, who are not. If you are young, you still want to do things that can shape the rest of your life. If you are older or if you are sick like me, you might not care so much to shape the rest of your life as much as to live it and appreciate what you can. These differences between us are unavoidable and understandable. But it means that no matter how much some of you are there for me, I still feel alone. You do not know what it's like to be dying, and you probably can't know, until it happens to you. (Chung 2017)

Chung's description and the study of the cancer patients mentioned above bring into relief how the transformative experience of terminal illness and dying (in modern Western societies) differs from other kinds of transformative experience.[14] In illness and dying, one's future appears highly contracted and one feels alone and often powerless. Compare this sense of a shrunken future and the feeling of a loss of control with the sense of an open future and a feeling of agency in the typical (affluent, Western) transformative experience of becoming a parent, especially as this experience is described in modelling the problem of transformative choice (e.g. Paul 2014; 2015c). This problem is presented as one in which you have to choose between two open-ended, possible futures before you—whether to become a parent or to remain child-free—and as one in which you have control over which scenario to bring about. In the experience of dying, however, this sense of an open future vanishes along with the feeling of control. The transformative experience of dying thus calls into question precisely the sense of self that has been used to pose the problem of transformative choice—namely, the experience of the self as a rational, autonomous agent in control of its life and its future.

[14] See also Carel etal. (2016).

We can draw a general lesson from this point. Not only are many experiences transformative without being the result of choice, but also the fact that an experience is not the result of choice can be part of what makes it transformative.

Let me connect these observations to an ancient philosophical idea. As Amber Carpenter writes in her paper, "Metaphysical Suffering, Metaphysics as Therapy," Greek grammar contains "an insight into the human condition":

> In an irregular formation of the passive voice, *paschein,* 'to suffer', is the ordinary passive form of *poiein,* 'to do'. To suffer is to have something done to one, or happen to one. It is particularly defined through its contrast class: doing, being active, and especially contrasted with being in control. Lack of control is suffering. (2011: 4)[15]

I take the insight here to be that suffering implies enduring or undergoing something you cannot control, but I do not think it is correct to state that lack of control *per se* is suffering. After all, many situations in which you lack control can, under the right circumstances, be experienced as joyful, thrilling, or liberating. Rather, it is the experience of wanting and trying to be in control and not being able to be in control that is suffering.

Trauma is a case in point, and one that has important connections to the experience of dying. Susan Brison, in her powerful book, *Aftermath: Violence and the Remaking of a Self*, in which she reflects philosophically on her own experience of being sexually assaulted and nearly murdered, writes, "One of the most serious harms of trauma is that of loss of control" (2002: 73). She defines a traumatic event as "one in which a person feels utterly helpless in the face of a force that is perceived to be life-threatening" (p. 39). She mentions "terror, loss of control, and intense fear of annihilation" as part of the immediate response to a traumatic event, and notes that the long-term effects "include the physiological responses of hypervigilance, heightened startle response, sleep disorders, and the more psychological, yet still involuntary, responses of depression, inability to concentrate, lack of interest in activities that used to give life meaning, and a sense of a foreshortened future" (pp. 39-40). The loss of control occurs not just when the traumatic event happens but also long afterward. People who suffer from post-traumatic stress disorder (PTSD) experience intrusive and emotionally overwhelming memories, heightened startle responses, and involuntary responses to things that previously provoked no response. "A trauma survivor suffers a loss of control not only over herself, but also over her environment, and this, in turn, can lead to a constriction of the boundaries of her will ... Some reactions that were under the will's command become involuntary and some desires that were once motivating can no longer be felt, let alone acted upon" (p. 60).

[15] I thank Jelena Markovic for bringing this paper to my attention, and I am indebted to her use of it in Jelena Markovic, "How to Die Before You Die: The Transformative Power of Meditation," presented

Loss of control is one reason that trauma transforms the self: "Such loss of control over oneself can explain, to a large extent, what a survivor means in saying, 'I am no longer myself '" (Brison 2002: 60). Part of what makes trauma a personally transformative experience is its resulting from an unwilled and unwanted loss of control. In Brison's case, the transformative experience of trauma involved a transformative experience of dying in two ways. First, she was the object of an attempted murder and experienced coming very close to death. Second, she describes her life afterward as if she were "experiencing things posthumously" (p. 8):

> When the inconceivable happens, one starts to doubt even the most mundane, realistic perceptions. Perhaps I'm not really here, I thought, perhaps I did die in that ravine. The line between life and death, once so clear and sustaining, now seemed carelessly drawn and easily erased.
>
> For the first several months after my attack, I led a spectral existence, not quite sure whether I had died and the world went on without me, or whether I was alive but in a totally alien world. (2002: 8-9)

Brison describes the experience of survival and recovery as one in which she had to let her former self die, making the trauma tantamount to a kind of death:

> I am not the same person who set off, singing, on that sunny Fourth of July in the French countryside. I left her in a rocky creek bed at the bottom of a ravine. I had to in order to survive. (2002: 21)

Brison reports that, in order to recover, the trauma survivor needs to be able to take control of herself, and that one of the important ways this happens is by constructing a narrative and telling it to an empathetic listener. The narrative enables the survivor "not only to integrate the traumatic episode into a life with a before and an after, but also to gain control over the occurrence of intrusive memories." Such control, "repeatedly exercised, leads to greater control over the memories themselves, making them less intrusive and giving them the kind of meaning that enables them to be integrated into the rest of life" (p. 54).

As a general matter, in situations in which you want and try to be in control but feel powerless to choose or to act, you experience suffering. Choosing and acting are rational activities, in the sense that they involve having and being able to give reasons, and being able to devise a meaningful narrative to make sense of what has happened or is happening to you and of what you do. So, engaging in meaningful activity, including constructing a narrative, can mitigate suffering.

Dying patients in hospice cite the loss of control, including the inability to preserve the meaningful narrative of a life with choice, action, and an open future, as one of

---

at "Crossing Over: An Interdisciplinary Conference on Death and Morbidity," York University, Toronto,

the principal causes of their suffering. Loss of control and the breakdown of meaning make the suffering an existential crisis. Palliative care specialists describe this kind of suffering as "intrinsic to the dying process" (Rattner and Berzoff 2016). Some specialists argue that it may not be possible to alleviate it, and that it can be a mistake to try to relieve it. Rather, the palliative care provider should acknowledge it with the patient, an approach called "sitting with suffering" (Rattner and Berzoff 2016).[16]

Sitting with suffering, for both the dying and those who care for them, implies a different approach to the felt loss of control from that of trying to regain control. The approach is not to try to control or manage the existential crisis, as one tries to control or manage physical pain, but rather to bear witness to the suffering. Sitting with suffering suggests a meditative posture—one that recognizes and observes the reality of suffering, without trying to assert control where there is ultimately none to be had. Brison, too, points out that during recovery from trauma, the effort to control your life through constructing a narrative can be taken too far and can hinder recovery, and that it is important to learn how to relinquish control and to let go (2002: 103, 115). Similarly, the task of the empathetic listener is not to control but rather to bear witness.

This kind of letting go of control is very different from the kind of control you are supposed to be exercising in making a transformative choice by trying to bring the future into line with your subjective values and preferences. Indeed, part of the challenge that the transformative experience of death poses is precisely that you may feel a pressing need to find a way to let go of that sense of self as controller. Nevertheless, it is crucial to realize that letting go of control in this way can be rational. It can stem from an accurate perception of how things are, namely, transient and essentially out-of-control, and from an acceptance in which you exercise your capacity for understanding. Being capable of such perception and acceptance is thus part of what it is to be well-equipped to approach death rationally.

Dying individuals and those who care for them often report that practicing this kind of perception and acceptance precipitates a transformation of the self and the emergence of new meaningful narratives. They report that suffering is ameliorated through this process, rather than through efforts at control. For example, in the cancer patients of the study mentioned earlier, losing the sense of self as a controlling agent initiated a phase of deep mourning, but the patients who were able to live through the intense mourning and let go of this sense of self unexpectedly experienced a deeper sense of self and belonging to a larger whole. One patient reported: "suddenly I heard the beating of my heart and I thought: yes, it is a piece of nature that is there. And then I felt myself being part of nature" (Yang et al. 2010: 60). This relinquishing of an "egocentered worldview in favor of a deep sense of being embedded in a larger whole" lessened these patients' loneliness and fear of death (p. 63).

February 17-18, 2017.

[16] See also Halifax (2008).

I have called attention to the shrunken sense of the future in the transformative experience of death. The retrospective perspective too differs from that of any other kind of transformative experience. You know that your time is coming to an end and that you have a last opportunity to review your whole life. Review and reminiscence, which can be spontaneous or deliberate, figure prominently in the dying experience (Butler 1963; Kellehear 2014: 149-67). In Kellehear's words:

> Any major life crisis can bring you to a point where you will interrogate the "past selves" in search of a discovery or rediscovery of the meaning of self, but dying will often do this too for many people because dying is the final roundup of all meaning about their life. People review their lives for another reason, too. Sometimes, it is a simple reacquaintance with the contents of memory. (2014: 159)

Reacquaintance with the contents of memory is especially poignant in people suffering from dementia. They cherish memories as a critical way to hold onto their sense of self as it slips away. Kellehear quotes these words from Thomas DeBaggio's memoir of living with Alzheimer's disease: "Even in this time of failing memory, I am happy to stay closeted in my mind and bring up broken memories to paw over" (p. 153).

Life review in the face of death occurs not just in the elderly but also in younger people and children. Kellehear describes three main forms the remembering takes. First, people can deliberately and selectively review important relationships and events. This kind of remembering serves self-understanding and anticipatory grieving for future loss. Second, people can experience an uncontrolled, non-selective, and panoramic life review, especially in traumatic situations. This kind of memory event is one of the elements used to classify near-death experiences (Greyson 1983). Third, people in prison or death camps remember and dwell on quotidian events of a better life, often in order to block out the terrible suffering of their present circumstances.

These kinds of remembering can retrospectively recast the meaning of earlier transformative choice points of your life, the meaning of the transformative decisions that you made, and the meaning of the transformative experiences that you had as a result. Imagine recalling at death your decision to become a parent, or to join the armed forces and go to war, or to abandon academic life and become a farmer, or to become a monk or a nun, or to renounce your monastic vows and return to secular life. Thus, another way that death is the ultimate transformative experience is that it can serve as an ultimate meta-perspective from which to assess the value and meaning of every other experience you have had, including especially those resulting from the transformative decisions of your life.

In his summary of how the experience of death is transformative, Kellehear distinguishes between the three phases of early dying, late dying, and the moment of death. Throughout the early and late dying phases, physical distress, mental suffering, and existential crisis mix with new insights and fresh perceptions. Early in the dying process,

> people are already noticing the world is different for them. They see the world afresh; their perceptions are rejuvenated; they start to notice things they did not see before in their environment. These dying people's altered perceptions provide them a new appreciation of life and what it is offering them day to day. (2014: 191)

Gaining these new experiences is epistemically and personally transformative. They stem from the growing realization of your impending dissolution, and deeply change how you experience your self. Kellehear states that the "most fundamental observation" about the dying experience is that "one leaves one's former self to become a new self or to integrate a new sense of self" (2014: 205). The transformation reorganizes old values and preferences, and brings new ones. As dying progresses, such changes increase, so that people come to expect them:

> [M]any dying people are themselves amazed at the how the world around them seems so different and how their inner life seems so in the throes of transformation that ever newer, more novel, ever strange, and foreign experiences seem not only to be possible but even very likely to them. Many dying people at the center of these new perceptions and changes come to expect more alterations to their inner realities and experiences. (Kellehear 2014: 194)

These alterations become pronounced in the late phase of dying and at the moment of death. Prolonged experiences of pain and distress can transition to feelings of peace, calm, and serenity (Kellehear 2014: 200). About one-third of dying people have deathbed visions, though their prevalence is lower in modern Western societies where opiates, which apparently suppress these visions, are routinely used in caring for the dying. Palliative care medicine is taking a new interest in deathbed visions and dreams as transformative experiences in their own right (Kerr et al. 2014). They do not seem to be predicted by diagnosis, religion, ethnicity, class, gender, or age, and are generally described as positive (Kellehear 2014: 196-7). Hospice workers also report that dying people often confront and reconcile themselves to what they take to be deep truths about themselves revealed in their final moments, and that they discover new truths.

We can relate this idea of death as revelatory to Paul's concept of "revelation," which she uses as a way to deal with the problem of transformative choice (Paul 2014). Recall that making a transformative choice requires you to imagine what the new transformative experience will be like and to determine the subjective value it will have for you; but you cannot do this before you have the experience, and therefore you lack the rational means to choose. Paul proposes that we reframe how we think about transformative choices by seeing them as presenting us with a choice between either embracing or avoiding what she calls "revelation." You can either choose to have the new experience for the sake of what it reveals to you, including new subjective values

and the discovery of a new sense of self, or you can choose to forgo the revelation and affirm your current life.

In the case of death, however, you cannot choose not to undergo the transformation. What you may be able to choose, to some extent, is how you undergo it. You may be able to choose to meet the transformation and what it may reveal with acceptance instead of trying to push it away, or you may choose to resist and fight it until the very end. Understanding and being prepared for these choices and possibilities is also part of what it means to be well-equipped to approach death rationally.

Here, "rationally" means being able to exercise your capacities for observation, understanding, choice, and letting go. It does not mean the limited, decision- theoretic sense of instrumental rationality, which is the kind of rationality usually emphasized in the philosophical literature on transformative experience. Indeed, one of the benefits of reflecting on what it might mean to be better or worse equipped to approach death rationally is that such reflection can help us to see how limited the instrumental sense of rationality may be for our understanding of transformative experiences altogether.

4. Death as Existentially Transformative

Death is not just epistemically and personally transformative; it is existentially transformative.[17] Death constitutes your finitude, and the knowledge that you will die and the transformative experience of dying reveal your finitude in a unique and all-encompassing way.

Any transformative situation confronts you with your finitude—with your having to go down the path of one transformative course of action rather than another. "To be finite, in fact, is to choose oneself—that is, to make known to oneself what one is by projecting oneself toward one possibility to the exclusion of others" (Sartre 1956: *698).[18] Sartre rejects what he sees as Heidegger's "strict identification of death and finitude" (p. 698). Being immortal, in the sense of having an unending lifespan— as in the Makropolous case discussed by Bernard Williams (1973: 82-100) or Paul's (2014: 1-4) example of becoming a vampire—would not remove your finitude, for you would still find yourself confronted with bounded choices and courses of action, even if there were an endless series of them. Nevertheless, immortality would radically change your finitude, and this fact is enough to show that death constitutes your finitude, even if it is not strictly identical with it.

Following Heidegger, we can say that your death is not just your final and inevitable transformation, but also the one that fully individuates you, and that ultimately no one can take away from you or stand in as a substitute for you (2010 [1927]: 227-56). This

[17] Carel et al. (2016) use the term "existentially transformative" to describe illness experience.

[18] Sartre (1956). I have slightly modified Barnes's translation, using "possibility" instead of her "possible." The French reads: "Etre fini, en effet, c'est se choisir, c'est-a-dire se faire annoncer ce qu'on est en se projetant vers un possible, a l'exclusion des autres" (Sartre 1943: 159).

existential characteristic of being "ownmost" enables you to take up a meta-perspective from which you can view any other of your possible transformations, including especially the actual transformative experiences of your life that you may revisit retrospectively when you are dying. Earlier I said that the transformative experience of death can retrospectively recast the meaning of prior experiences. I can now expand upon this point in the following way. The meaning of particular events in your life and of your life as a whole depends on that which individuates you as the finite being that you are; such individuation essentially includes your own death; and how you choose to live in the light of the realization of your own "being-toward-death," including how you mentally prepare yourself for death, determines how well- or poorly-equipped you are to approach death rationally and authentically.

Heidegger's conception of "authentic being-toward-death," however, is radically defective and deficient. He contrasts the public and inauthentic de-individualizing and depersonalizing of death, whereby we all think "one dies" or "everyone dies," with the authentic awareness of one's own death as one's "ownmost possibility."[19] But he also asserts that this "ownmost possibility" is "nonrelational," by which he means that all your relations to others have no place in it (2010: 239). Your experience of the deaths of others, as well as your relations to others as you face and undergo your own death, are absent from what Heidegger nonetheless calls "the full existential and ontological concept of death" (p. 239) As a result of making one's own self primary and one's relations to others secondary, Heidegger distorts the human experience of death by severing its links to grief and love. (Being and Time says nothing about grief and love.)

I do not deny that one's own death is one's "ownmost possibility." I deny that it is "non-relational." Death is relational because it is interpersonal and intersubjective. Chung illustrates the point in the concluding words of his essay, "Is Dying a Transformative Experience?"

> I'm not the only one this illness has transformed. It has turned my wife into someone whose husband is dying. And it will eventually turn my wife into a widow. I do not and cannot know what it's like to watch your husband or wife suffer through this illness, and I will not know and cannot know what it's like to lose the love of my life. I can only imagine what it's like for her now and what it will be like for her then, and she can only imagine what it's like for me—and we both know that such imaginings can only give us the barest outline of what our experiences are like. There are two of us transformed by this illness, but in different ways. My wife and I are both left a little alone by this illness, left incompletely understood, even by each other.

[19] To my mind, Tolstoy, in *The Death of Ivan Ilyich,* depicts this contrast better and more vividly. See Tolstoy (2006 [1886]).

> But despite that, we are here for each other. And even if none of my friends and family really understand what it's like to be dying or what it's like to have a husband who's dying, they are here for us—unwavering and stalwart. Now that life seems so fragile and short to me, all this seems like a wonder. (Chung 2017)

A better perspective than Heidegger's on authentic being-toward-death, one that keeps death, grief, and love together, comes from the early Indian Buddhist story known as "Kisa Gotami and the Mustard Seed."[20] The story is found in a commentary on a poem from the *Therigatha,* a collection of short poems in the Pali language by and about the first Buddhist women (Hallisey 2015). These women were called *theri,* "elders," because they were the senior, ordained, female monastics. The *gatha,* "verses," are their poems.

Our poem is said to be chanted by Kisa Gotami. The contemporary American poet Anne Waldman renders the poem as follows:

> Excellent to have wise, noble friends
> One should know a few things
> It helps your pain
> But one should understand how pain arises
> how it ceases
> (the Eightfold Path, the Four Noble Truths)
> Mark the sorrow, mark how it comes
> Being a woman is painful
> Miserable sharing a home with hostile wives
> Some cut their own throats
> More squeamish women take poison
> ...
> But I survived
> quenched desire
> Saw the teaching as a mirror
> held up to show me my crazy mind
> Now healed
> The poison darts extracted from my heart
> All this done
> and done by me
> The theri Kisa-Gotami
> saw herself in the mirror
> and witnessed these things. (Shelling and Waldman 1996: 69-70)[21]

[20] See Buswell and Lopez (2014: 437).

[21] For a more scholarly translation, see Hallisey (2015: 110-15).

The commentary on this poem tells Kisa Gotami's story. She is born into a poor family and named Gotami but called "Kisa" ("lean" or "emaciated") because she is so thin. She has the good fortune to marry into a wealthy family, but she receives no respect and is badly treated until she gives birth to a son. Her happiness and good fortune, however, are short-lived, for her son dies when he becomes old enough to run around and play. Mad with grief and having lost her family status, she refuses to part with her child's body and wanders everywhere, carrying the body on her hip, searching for medicine to bring him back to life. Everywhere she goes she is mocked and driven away. Finally, a kind man takes pity on her and directs her to the Buddha. She begs the Buddha for medicine to revive her son, and he tells her to fetch a mustard seed from a house in which no one has died. She goes from house to house, searching frantically, but everywhere she goes she cannot find a single house in which no one has died. (This is the origin of the aphorism, "The living are few, but the dead are many," which Paul Carus used in his 1894 rendition of the story: Carus 1894/ 1915 [1894]: 209-13.) Eventually it dawns on Kisa Gotami that death is inevitable and common to all, and that everyone experiences grief when loved ones die. She lays her son's body to rest in the charnel ground, and speaks this verse:

> No village law is this, no city law,
> No law for this clan, or that alone;
> For the whole world—and for the gods too—
> This is the law: All is impermanent. (Elbaum Jootla 1994)

She returns to the Buddha and he asks her whether she has obtained the mustard seed. She answers that the matter of the mustard seed is finished, and the Buddha utters this verse:

> A person with a mind that clings,
> Deranged, to sons or possessions,
> Is swept away by death that comes
> —Like mighty flood to sleeping town. (Oldenzki 2005)

She asks to be admitted to the monastic order, is ordained, and eventually becomes an arahant, a person who has attained liberation. The Buddha proclaims her the foremost of those who wear coarse robes, and this status is said to be the attainment of an aspiration and a vow that she made in a previous existence when she witnessed an earlier Buddha bestow the same status on another woman monastic. The later tradition recognizes her for being outstanding in ascetic practices (Analayo 2014a: 97-115).

Many things could be said about how we today may find things in this story that resonate with our own sensibilities—the recognition of the experience of being a woman and the esteem given to the woman's voice, the rejection of mainstream patriarchy in favor of a counterculture sustained by a deeper understanding and called to a higher

purpose, and the use of poetry to express personally transformative experiences of realization. In addition, I find a deep and moving insight into our being-toward-death, an insight that can help us to see the failure of Heidegger's account.

Whereas Heidegger, in his analysis of being-toward-death, moves from the anonymous public world to the individual, the Kisa Gotami story moves from the individual to the community, and then back to the individual and a new sense of community. Heidegger takes inauthentic being-toward-death, in the form of the anonymous, public thought "one dies," and opposes it to authentic being-toward- death, described as the realization of my own death as my "ownmost, nonrelational" possibility. Our relations to others finds no place in his conception of our being- toward-death. Reading Heidegger refracted through the Kisa Gotami story reveals the deficiency of his account: Love and grief make possible authentic being-toward- death, but Heidegger says nothing about them. His famous statement, "Insofar as it 'is', death is always essentially my own" (2010 [1927]: 223), is faulty. I agree with Simon Critchley (2009) that Heidegger's "conception of death is both false and morally pernicious" to the extent that it makes the deaths of others secondary to my own death. As Critchley writes,

> On the contrary ... death comes into our world through the deaths of others, whether as close as a parent, partner, or child or as far as the unknown victim of famine or war. The relation to death is not first and foremost my own fear for my own demise, but my sense of being undone by the experience of grief and mourning. (2009)

The Kisa Gotami story presents this truth. The story starts from the individual experience of love, loss, and grief, and moves to an empathetic understanding of grief and death as shared by all. The realization is existentially transformative: It transforms her grief, her understanding of who and what she is, and her understanding of the whole human world (as well as her understanding of all sentient beings, including gods). She has an existentially transformative insight into the finitude of sentient existence. The insight transforms her values, preferences, and emotions, so that she lets go of her former identity and community—a scorned and grieving village wife—leaving it in the charnel ground with her dead child's body, and she joins a new community, one devoted to the practice of the mindfulness of death and the quest for liberation.[22]

5. Conclusions

This chapter began with the thought that death is the ultimate transformative experience. Now, at the end of the chapter, we have been brought to see that there is an important sense in which death is not the only ultimate transformative experience.

[22] For translations of the early scriptural presentations of the mindfulness of death, see Analayo (2016).

Although death alone may be ultimate in the sense of being final, it is not alone in being ultimate in the sense of having fundamental significance and in having the power to transform everything else. Love and grief are also ultimate transformative experiences. So, too, is liberation, from the perspective of Buddhist philosophy (as well as Indian religion and philosophy in general). All provide metaperspectives from which to view everything else in one's life.

This chapter also began with how Paul poses the problem of transformative experience, which is by arguing that you cannot grasp a personally transformative experience until you undergo it and learn how it changes you. Without experiencing dying, you cannot fully comprehend what it is like to die. There is no way to gainsay this fact. At the same time, we have also been brought to see that there may be ways to witness death that can give us insight into what it is like to die. Being a witness in this sense does not mean being a mere observer. It means being with the dying person and bearing witness to that person's experience. It means being empathetically present to the person's own unique experience of dying, and experiencing the loss of this person from your own life and from the world altogether. One reason we have so little insight into the experience of death may be that our culture resists being with the dying and bearing witness to the reality of death.

Let me end this chapter by reminding you of the question of what it might mean for us to be able to approach death rationally and authentically, and by giving the last word to Kisa Gotami. While she is meditating, Mara, the Indian demon associated with death, tries to distract her with his terrifying presence, just as he tried to distract Siddhartha Gautama from his meditation on the evening of his awakening when he became the Buddha. Mara sings to Kisa Gotami this verse:

> Why are you, having lost your child,
> Weeping and with sad and worried face
> sitting alone under a tree?
> Are you searching for a man?

Kisa Gotami answers:

> Without limit are the sons,
> who all have died and been lost.
> This, then, is the end of men [for me].
> I have gone beyond [the attraction of] men's external appearance.
> Not troubled, not sad or worried,
> I have done what should be done in the Buddha's dispensation. Separated from all craving and *dukkha*,
> having entirely relinquished the darkness [of ignorance],
> I have realized cessation,
> I dwell in peace and at ease with influxes eradicated.

I recognize you, evil Mara,
now make yourself disappear and go! (Analayo 2014b: 123-4)

Mara, sad, afflicted, and annoyed, thinks, "Kisa Gotami knows me and my intentions," and vanishes.[23]

References

Analayo, B. 2014a. "Outstanding Bhikkhunis in the *Ekottarika-agama."* In A. Collett (ed.), *Women in Early Buddhism: Comparative Textual Studies,* 97-115. Oxford: Oxford University Press.

Analayo, B. 2014b. "Denying Mara: Bhikkhunis in the SaKyukta-agama." In A. Collett (ed.), *Women in Early Buddhism: Comparative Textual Studies,* 123-4. Oxford: Oxford University Press.

Analayo, B. 2016. *Mindfully Facing Disease and Death: Compassionate Advice from Early Buddhist Texts.* Cambridge: Windhorse Press.

Barnes, E. 2015. "Social Identities and Transformative Experience." *Res Philosophica* 92: 171-87.

Bennett, G. 2014. "Goodbye and Goodluck!" <`http://www.deadatnoon.com/index.html`>
Borjigin, J., U. Lee, T. Liu, D. Pal, S. Huff, D. Klarr, ... and G. A. Mashour. 2013. "Surge of

Neurophysiological Coherence and Connectivity in the Dying Brain." *Proceedings of the National Academy of Sciences USA* 110: 14432-7.

Brison, S. J. 2002. *Aftermath: Violence and the Remaking of a Self.* Princeton, NJ: Princeton University Press.

Buswell, R. E., Jr., and D. S. Lopez, Jr. (eds) 2014. "Kisa Gotami." In *The Princeton Dictionary of Buddhism.* Princeton, NJ: Princeton University Press.

Butler, R. N. 1963. "The Life Review: An Interpretation of Reminiscence in the Aged." *Psychiatry* 26: 65-76.

Campbell, J. 2015. "L. A. Paul's *Transformative Experience." Philosophy and Phenomenological Research* 91: 787-93.

Carel, H., I. J. Kidd, and R. Pettigrew. 2016. "Illness as Transformative Experience." *The Lancet* 388: 1152-3.

[23] I wish to thank The Experience Project (the-experience-project.org) and the Templeton Foundation for a fellowship that supported the writing of this chapter. For helpful comments, I wish to thank Bhikkhu Analayo, Anthony Bruno, Richard Jaffe, Troy Jollimore, Enoch Lambert, Jelena Markovic (whose research assistance was also invaluable), Laurie Paul, Geoffrey Sayre-McCord, Robert Sharf, John Schwenkler, Sean Smith, Jenny Windt, and the graduate students in my 2017 seminar on transformative experience and the philosophy of death: Katherine Cheng, Kyle Da Silva, Jasper Heaton, Andrew Jones, Matthew Kinakin, Jason Leslie, Phyllis Pearson, Matthew Smithdeal, Richard Wu, and Steven Zhao. This chapter derives from papers presented at the American Philosophical Association Pa-

Carpenter, A. D. 2011. "Metaphysical Suffering, Metaphysics as Therapy." In B. Hogue and A. Sugiyama (eds), *Making Sense of Suffering: Theory, Practice, Representation,* 3-10. Oxford: Inter-Disciplinary Press.

Carus, P. 1915 [1894]. *The Gospel of the Buddha.* Chicago : Open Court.

Chung, A. K. 2017 (August 23). "Is Dying a Transformative Experience?" <https://professorkenchung.wordpress.com/2017/08/23/is-dying-a-transformative-experience/>

Critchley, S. 2009 (July 13). "Being and Time: Part 6." *The Guardian.* <https://www.theguardian.com/commentisfree/belief/2009/jul/13/heidegger-being-time>

DeGrazia, D. 2017. "The Definition of Death." In *The Stanford Encyclopedia of Philosophy.*
<https://plato.stanford.edu/entries/death-definition/>

Doughty, C. 2015. "The Sacred Task of Caring for the Dead Should Be a Normal Part of Life." *The Guardian* (April 13). <https://www.theguardian.com/commentisfree/2015/apr/13/ sacred-task-caring-for-dead-home-funeral-grief>

Dowling Singh, K. 1998. *The Grace in Dying.* New York: HarperCollins.

Edwards, P. 1969. "Existentialism and Death: A Survey of Some Confusions and Absurdities." In S. Morgenbesser, P. Suppes, and M. White (eds), *Philosophy, Science and Method: Essays in Honor of Ernest Nagel,* 473-505. New York: St. Martin's Press.

Elbaum Jootla, S. 1994. "Inspiration from Enlightened Nuns." <https://www.accesstoinsight.org/lib/authors/jootla/wheel349.html>

Greyson, B. 1983. "The Near-Death Experience Scale: Construction, Reliability, and Validity." *Journal of Nervous and Mental Disease* 71: 369-75.

Hacking, I. 1996. "The Looping Effects of Human Kinds." In D. Sperber, D. Premack, and A. J. Premack (eds), *Causal Cognition: A Multidisciplinary Debate,* 351-83. Oxford: Oxford University Press.

Hadot, P. 1995. *Philosophy as a Way of Life: Spiritual Exercises from Socrates to Foucault,* ed. A. Davidson. Malden, MA: Blackwell.

Halifax, J. 2008. *Being with Dying: Cultivating Compassion and Fearlessness in the Face of Death.* Boston, MA: Shambhala.

Hallisey, C. (trans.). 2015. *Therigatha: Poems of the First Buddhist Women.* Cambridge, MA: Harvard University Press.

Harman, E. 2015. "Transformative Experiences and Reliance on Moral Testimony." *Res Philosophica* 92: 323-9.

Heidegger, M. 2010 [1927]. *Being and Time,* trans. J. Stambaugh. Albany, NY: State University of New York Press.

Henrich, J., S. J. Heine, and A. Norenzayan. 2010. "The Weirdest People in the World?" *Behavioral and Brain Sciences* 33: 61-135.

Hospice Patients Alliance. n.d. "Signs and Symptoms of Approaching Death." <https://www. hospicepatients.org/hospic60.html>

Kellehear, A. 2014. *The Inner Life of the Dying Person.* New York: Columbia University Press.
Kerr, C. W., J. P. Donnelly, J. P. Wright, S. M. Kuszcack, A. Banas, P. C. Grant, and D. L. Luczkiewicz. 2014. "End of Life Dreams and Visions: A Longitudinal Study of Hospice Patients' Experiences." *Journal of Palliative Medicine* 17: 296-303.
Kroeger, D., B. Florea, and F. Amzica. 2013. "Human Brain Activity Patterns Beyond the Isoelectric Line of Extreme Deep Coma." *PLOS One* 8: e75257.
Kuhl, D. 2002. *What Dying People Want.* Toronto: Anchor Canada.
Laureys, S. 2005. "Death, Unconsciousness, and the Brain." *Nature Reviews Neuroscience* 6: 899-909.
Luper, S. 2016. "Death." In *The Stanford Encyclopedia of Philosophy.* <https:// plato.stanford. edu/archives/sum2016/entries/death/>
Morison, R. S. 1971. "Death: Process or Event?" *Science* 173: 694-8.
Nagel, T. 1986. *The View from Nowhere.* Oxford: Oxford University Press.
Nair-Collins, M. 2015. "Taking Science Seriously in the Debate on Death and Organ Transplantation." *Hastings Center Report* 45: 38-48.
Norton, L., R. M. Gibson, T. Gofton, and C. Benson. 2017. "Electroencephalographic Recordings During Withdrawal of Life-Sustaining Therapy Until 20 Minutes After Declaration of Death." *Canadian Journal of Neurological Sciences* 44: 139-45.
Not Nurse Ratched. 2016 (April 21). "Why Do We Call It 'Actively Dying'?" <http:/ / notratched.net/nnr/2016/4/21/why-do-we-call-it-actively-dying>
Nutik Zitter, J. 2017. "First, Sex Ed. Then Death Ed." *New York Times* (February 18).
Oldenzki, A. 2005. "Skinny Gotami and the Mustard Seed." <https://www.accesstoinsight.org/ noncanon/comy/thiga-10-01 -aoo.html>
Parnia, S. 2014. "Death and Consciousness: An Overview of the Mental and Cognitive Experience of Death." *New York Academy of Sciences* 1330: 75-93.
Paul, L. A. 2014. *Transformative Experience.* Oxford: Oxford University Press.
Paul, L. A. 2015a. "Transformative Choice: Discussion and Replies." *Res Philosophica* 92: 473-545.
Paul, L. A. 2015b. *"Transformative Experience:* Replies to Pettigrew, Barnes and Campbell." *Philosophy and Phenomenological Research* 91: 794-813.
Paul, L. A. 2015c. "What You Can't Expect When You're Expecting." *Res Philosophica* 92: 1-23.
Rattner, M., and J. Berzoff. 2016. "Rethinking Suffering: Allowing for Suffering That Is Intrinsic at End of Life." *Journal of Social Work in End-of-Life and Palliative Care* 12: 240-58.
Sartre, J.-P. 1956 [1943]. *Being and Nothingness,* trans. H. E. Barnes. New York: Washington Square Press.

cific Division Pre-Conference, "Themes in Transformative Experience," April 11, 2017; the Philosophy Graduate Conference at the University of Toronto, May 8-9, 2017; and the Chapel Hill Workshop on Transformative Experience, May 11, 2017.

Shelling, A., and A. Waldman (trans.) 1996. *Songs of the Sons and Daughters of Buddha.* Boston: Shambhala.

Thompson, E. 2015. *Waking, Dreaming, Being: Self and Consciousness in Neuroscience, Meditation, and Philosophy.* New York: Columbia University Press.

Tolstoy, L. 2006 [1886]. *The Death of Ivan Ilyich,* trans. A. Briggs. London: Penguin.

Valberg, J. J. 2007. *Dream, Death, and the Self.* Princeton, NJ: Princeton University Press.

Warraich, H. 2017. *Modern Death: How Medicine Changed the End of Life.* New York: St. Martin's Press.

Warren, J. 2004. *Facing Death: Epicurus and His Critics.* Oxford: Oxford University Press.

Williams, B. 1973. "The Makropolous Case: Reflections on the Tedium of Immortality." In B. Williams (ed.), *Problems of the Self,* 82-100. Cambridge: Cambridge University Press.

Yang, W., T. Staps, and E. Hijmans. 2010. "Existential Crisis and the Awareness of Dying: The Role of Meaning and Spirituality." *Omega* 61: 53-69.

Index

The Ted K Archive

John Schwenkler & Enoch Lambert
Becoming Someone New
Essays on Transformative Experience, Choice, and Change
2020

ISBN 0198823738, 9780198823735

Oxford University Press

www.thetedkarchive.com