Technological Danger

Retraice

Jan 12, 2023

Contents

Part 1: Uncertainty, Fear and Consent

Transcript

OK. And we're live retrace 113 for January 11th, 2023. We're going to talk about technological danger. This is sort. Of a setup segment. We're going to get through a lot of stuff that we're going to come back to. Let's get started.

Prediction: freedom is going to decrease

We'll start with the prediction. The Freedom, security safety trade off will continue to shift towards safety over the next 20 years. Between 2023 and 2032 inclusive, you'll continue to be asked, told and nudged into giving up freedom in exchange for safety, which is about unintentional danger. In addition to security, which is about intentional danger, and I get that distinction from Bruce Schneider and one of his early books, other people to consider on this in addition to Schneider, I guess Norbert Weiner, human use of human beings, Russell Impact this is science. In society, we've talked about George Dyson's books, Samuel Butler, Kurzweil's spiritual machines, Ted Kaczynski technological slavery, and then the couple of Nick Bostrom papers we've been talking about recently. The what are they called again? Information hazards and then the vulnerable world hypothesis. OK, so let's just things to think. About don't worry that that's a lot to cover. We're just talking about the prediction right now. Next 20 years is the shift is going to continue the trade off toward safety and of course, security.

Decisions: two problems of life and the problem of death

OK, there's a decision. A couple of them actually that follow from that. We've talked about the 2 problems of life and the problem of death before we started in R27 and we we that was one of the early world models or I think it was world Model 4 and then world Model 5331. Integrated it so the two problems of life are how to change the world and and how to change oneself. That part of the world that that we call ourselves. Those are the 2 problems of life recurring decisions in a sense, decision problems, recurring decision problems, and the problem of death is that dead things rarely become alive, whereas alive things regularly become dead. What do you do about that? That's a decision problem. It's not just an observation, it's a decision problem. You're in the game, you have skin in the game because presumably you don't

want to become dead. And yet that's happening all around you. And it's very hard to. The natural world to life. OK, so decisions. So far we've got prediction. The basic prediction, which is this shift towards safety and security, the trade off from freedom to safety and security giving up freedom goes down safety and security. Supposedly or or. Hopefully we'll go. Up and then the decision. One way of thinking about the the recurring decisions that will lead to that is, is the 2 problems of life. How do you change the world versus how do you change yourself? So if you change the world, you might make it more secure if you change yourself you. Just give up. Some security give up some freedom. OK. Problem of death is you. Why? Why do you care about safety and security? ultimately, I mean, you want to keep your stuff, you want to keep your limbs. So ultimately you want to keep breathing. OK so far so good.

Uncertainty

Let's talk about uncertainty. We just don't know that much about the future. We talked within the confines of our memories and instincts. Obviously, in a sense, your whole world is your memories or your whole sense of the world, or your whole, your whole sense of self or your. Whole presentation of. Self is is the better luck of it is in your in your memories. But it's also in your bodies memories in the form of genes. So we know the world through written history, more or less the last 5000 years of. Of Earth, if it's some form of it, obviously we have no real way of being sure about the history books, but there are a lot of things that have to come together in the wrong way for the history books to be totally unreliable. History is not bunk, but it's also as Henry Ford said. But it's also not. It's also not crystal clear our bodies know the earth about say, 2 billion years back or so in the form of genes more or less. Life arose at that point on Earth more or less 2 billion years. It's about 4 billion years old. Life, I think, started around 2 billion years ago. So. So in a sense. Our bodies know. Have some certainty about. The way the world works in the form of our genes. But if we're talking about survival. The parts of our bodies that know are are genes and the parts that would survive of the genes, and they can survive in. Other animals we. Share most of our DNA with the rest of the living world. And so we don't have to survive for that part. That cleverer part of the world that has a sense of what the world is really like. What the broader world is really like we don't have to... We humans don't have to survive for that part to survive. So we shouldn't assume that we're going to survive.

And there is hope in the form of controlling the environment to protect ourselves via technology, although we'll see that there's an irony in that controlling the environment via technology. But we also like to control the environment just to just to enjoy ourselves, not to. Preserve survival. Obviously, survival comes first in theory, but only if you can move quickly enough and depending on how quickly things are moving, you might or might not move quickly enough.

So we should think about technology. Causing danger, but also that it is the solution to danger. It's the solution to safety and security. It's not the solution, but it's it's it's an obvious category of solutions. Let's let's say it's, let's say that.

Beliefs and feelings

Let's talk about beliefs and feelings. So I just put together a couple of double s and a few double's here. So you believe there's a cure? You have a disease, you believe there's a cure. The feeling you have, you feel this hope. If you believe there's no cure for your disease. Fear among many other things, of course. If you believe there's a spaceship, you might start to feel excited. If you have any desire to ride in the spaceship, the younger you are, the more likely right. If if if your home town is the same. If you believe your hometown is the same, then you might feel longing homesickness. If you believe your hometown is not the same, you might feel sadness for not being able to go back to it. Uh, if you believe she loves you happiness if she. If you believe she hates you misery and if you believe she. Picks her. Knows disgust.

Just establishing the connections. The connection between beliefs and feelings, OK, so the subtitle for the notes or beliefs and the feelings they cause. Determine what chances we take, but possibilities don't. Care about our beliefs. That's just a little bit of the. Of the relationship between beliefs and feelings.

Chances

What about chances? Even getting out of bed as something as simple and and and every day as getting out of bed or not getting out of bed is somewhat risky. I don't have an example of this, but I'm sure if I looked I could find some poor *** ** * ***** who got out of bed and died immediately. I mean, you just even from just falling. I mean, actually, the older you get, the more likely it is that a fall is going to be your demise because you're more frail and you're falling. More falling is really dangerous. We're upright. It's just had one of my family members. Yesterday, but she was fine. But she was young. Falling's no good. Were these sticks with these walking stick bugs? And all the valuable stuff at. The top because I don't. Know what the evolutionary history is of of the. Location of our brains, but. Anyway, it falls the farthest. And the hardest when we fall. So getting out of bed is is risky. Staying in bed is risky. You're taking a chance if you're bedridden in a hospital of getting bed sores and all these terrible things that can lead to your demise as well. Whether or not you're old, you can. You can definitely be doomed by staying in bed, not getting out of bed. And and we do or don't get out of bed based on based on beliefs and instincts there's an instinct to sort. Of get up when? You wake up, but it's also about beliefs. I mean, what's that? You. It's easier to think of that saying that he's so depressed. He doesn't

want to get out of bed in the morning, right? Why is that? Instinct. Well, maybe. But it's. Also, belief that things are going to get better. It's nothing you can do. You get out of bed. There's nothing you can do. It's not your life is not going to get better, OK? It's a belief that. That that controls a chance that you take. What's the chance just getting out of bed, getting out of it? OK, if getting out of bed can be chancy, then anything can be chancy. And I just want to insert here something we've talked about in the past that was this would have been on November 27th that it's 1127, doesn't matter, OK. It's in the notes. The the radical economist von Meese has had a theory of human action. He wrote a whole treatise on it, and he said the three preconditions for human action. So, for example, getting out of bed or doing anything, giving up some freedom or not giving up some freedom, the three preconditions for action, our uneasiness with. Present number one, an image of a desirable future is number 2, and then the belief the expectation that action has the power to yield the image. Become what is it because humans have enough picking low hanging fruit, enough time, blah blah. OK, shouldn't put that in there, that's. From something else, OK, so. And I think I should. I should have said at the top that like I don't. Have an opinion about this freedom versus. Safety and security or freedom versus solutions to danger. I don't. I really don't like when. Most of the time, when I hear people talk about the trade off, it's always between freedom and security. They never talk about safety, but safety comes up, obviously in relation to technology and artificial intelligence safety, but also synthetic biology safety, that thing. I don't. I don't have any strong instincts on this. And when I think about losing freedom, I want to keep my. Freedom. When I think about losing my life. Losing quality of life in a profound way. I want to keep that and I'll give up some security. I'll give up some freedom for that safety and security. So you're getting. You're getting about as unbiased a person as you can on this, I think.

Possibilities

Let's talk about possibilities. If we're talking about uncertainty, we're talking about what's possible. We don't know that much about the future. We we know that the 5000 year Earth history, and we know the 2 billion year history and some sense deep in. Bones. What? What possibilities can we imagine? Well, like it's, you could go wild with this, but let's you know a radically good future based on technology. Let's imagine a cure for everything. You can imagine a radically bad future based on technology, synthetic plague that kills everybody, or worse than kills makes it misery. It's us, but doesn't kill us. You can imagine. So these are this is those are because like we can say that they're caused by technology although technology. Most of us. Would still argue that technology is being done by humans, So what about a radically good future because of humans? Imagine doctors inventing cures, and then we can imagine a radically bad future because of humans. Doctors inventing the synthetic plague. You can go on and on like this. The point is to distinguish the possibilities that you can

imagine from. From as what they are, which is just possible futures and then. And and imagine the Venn diagram of of like physically possible futures that is much bigger than your little. Your little dot inside of it that that represents which what you can imagine and what you can imagine is not it doesn't line up one to one with what other people can imagine. I'm sure there people out there who can imagine really, I mean the science fiction authors, right, like that's their job. But there are also people who know things that enable them to imagine things. That we the outsiders can't.

If you believe x, do you consent to y?

OK, so now let's just play a little game to wrap this up. If you believe X, do you consent to Y? OK, so if you believe that no one has privacy, do you consent to privacy invasion? Invasion is kind of loaded term, but it's you. Know if if you. Would you? Accept the NSA spying, or would you accept that every. Web-based company or web-based transaction involves cookies that are that destroy privacy in a sense. Do you accept it? do you accept that phones are like just tracking devices that happen to make phone calls? If you if you believe that you were the only person without that privacy that everyone else had privacy, would you still accept those? Things OK, so if you believe one thing, you consent to another thing, we all consent to the things I just listed more or less. I mean, you're presumably not you don't have a browser extension and and a million other things that protect you from the things I just. Described you don't have a. A jitterbug phone, right? You have a smartphone. OK, so you believe certain things and then you. Consent to other things. If you believe that entity XI should have used a different symbol here. If you believe that an entity is not malicious, do you consent to open interaction with that entity? If you believe that the NSA is not, let's say the IRS is not malicious, do you consent to open interaction with them? Telling them answering honestly, some people do. Some people don't. In the United States, so the UN, I don't know if they make it more international Google to make it more corporate, the NSA to make it more espionage. If you believe that these organizations. As a whole, or mostly or or completely are are not malicious. Do you consent to and answers? Yeah, you probably do. Right. You'd have to believe something about their malice and intentional or otherwise malice and intent or malice in effect, to to not openly interact with them. I think that I think that covers. Most of the the the refusals to openly interact OK. If you believe the vulnerable world hypothesis, do you consent to a global police state? We're going to talk about this. This comes from Boston's vulnerable world hypothesis paper from 2019, but the the, the, the short version of it is, if if technological. Development and this is the vulnerable world hypothesis I'm quoting Bostrom. Now, if technological development continues, then a set of capabilities will at some point be attained. That make the devastation of civilization extremely likely unless civilization sufficiently exists. Exists or exit? I'm sorry unless civilization sufficiently exits, the semi anarchic default condition and the semi anarchic default condition that he says we're in

now is limited capacity for preventive. Policing limited capacity for global governance and then diverse motivations. Those are the three things that roll up into the default and a semi anarchic condition. So what's a diverse motivate? What does that mean? Diverse motivations? There's a wide and recognizably human distribution of motives represented by a large population of actors at both the individual and state level. In particular, there are many actors motivated to a substantial degree by perceived self-interest, EG money, power, status, comfort and convenience, and there are some actors, the apocalyptic residual who would act in ways that destroy civilization even at high cost to them. Themselves. That's so that's his explanation of diverse motivation. So if you believe the vulnerable world hypothesis that it's some technological point of technological development, the devastating devastation of civilization is extremely likely, unless we do something about it. Do you consent to what, Bostrom. And and others would say is the logical thing to do, which is basically a global police state. We're going to talk later about. What that looks like he he describes it in detail. You're going to be wearing things around your neck. I'll give you a little bit of a teaser there, but for now that's it. OK, retrace 113 signing off.

Notes

Beliefs, and the feelings they cause, determine what chances we take; but possibilities don't care about our beliefs.

A prediction about safety, security and freedom; decisions about two problems of life and the problem of death; uncertainty, history, genes and survival machines; technology to control the environment of technology; beliefs and feelings; taking chances; prerequisites for action; imagining possibilities; beliefs that do or don't lead to consent; policing, governance and motivations.

Air date: Wednesday, 11th Jan. 2023, 10:00 PM Eastern/US.

Prediction: freedom is going to decrease

The freedom-security-safety tradeoff will continue to shift toward safety and security.

Over the next 20 years, 2023-2032, you'll continue to be asked, told, and nudged into giving up freedom in exchange for ***safety*** (which is about unintentional danger), in addition to **security** (which is about intentional danger).[1]

(**Side note:** We have no particular leaning, one way or another, about whether this will be a good or bad thing overall. Frame it one way, and we yearn for freedom; frame it another way, and we crave protection from doom.)

For more on this, consider:

- Wiener (1954);
- Russell (1952);
- Dyson (1997), Dyson (2020);
- Butler (1863);
- Kurzweil (1999);
- Kaczynski & Skrbina (2010);
- Bostrom (2011), Bostrom (2019).

[1] Schneier (2003) pp. 12, 52.

Decisions: two problems of life and the problem of death

First introduced in Re27 (Retraice (2022/10/23)) and integrated in Re31 (Retraice (2022/10/27)).

Two problems of **life**:

1. To change the world?
2. To change oneself (that part of the world)?

Problem of **death**:

1. Dead things rarely become alive, whereas alive things regularly become dead. What to do?

Uncertainty

We just don't know much about the future, but we talk and write within the confines of our memories and instincts.

We know the Earth-5k well via written history, and our bodies 'know', via genes, the Earth-2bya, about the time that replication and biology started. But the parts of our bodies that know it (genes, mechanisms shared with other animals), are what would reliably survive, not us. Most of our genes can survive in *other* survival machines, because we share so much DNA with other 2 creatures.[2]

But there is hope in controlling the environment to protect ourselves (vital technology), though we also like to enjoy ourselves (other technology). There is also irony in it, to the extent that technology itself is the force from which we may need to be protected.

Beliefs and feelings

- a cure, hope;
- no cure, fear;
- a spaceship, excitement;
- home is the same, longing;
- home is not the same, sadness;

[2] On creatures as gene (replicator) 'survival machines', see Dawkins (2016) pp. 24-25, 30.

- she loves me, happiness;
- she hates me, misery;
- she picks her nose, disgust.

Chances

Even getting out of bed—or not—is *somewhat* risky: undoubtedly some human somewhere has died by getting out of bed and falling; but people in hospitals have to get out of bed to avoid skin and motor problems.

We do or don't get out of bed based on instincts and beliefs.

Side note: von Mises' three prerequisites for human action:[3]

1. **Uneasiness** (with the present);
2. **An image** (of a desirable future);
3. The **belief** (expectation) that action has the power to yield the image. (**Side note**: technology in the form of AI is becoming more necessary to achieve desirable futures, because enough humans have been picking low-hanging fruit for enough time that most of the fruit is now high-hanging, where we can't reach without AI.)

Possibilities

- radically good future because of technology (cure for everything);
- radically bad future because of technology (synthetic plague);
- radically good future because of humans (doctors invent cure);
- radically bad future because of humans (doctors invent synthetic plague).

The important point is to remember the venn: there is a large space of possibilities, within which a small dot is what any individual human can imagine.

[3] von Mises (1949) pp. 13–14. See also Koch (2007) p. 144. See also Retraice (2022/11/27).

If you believe x, do you consent to y?

- no one has privacy, privacy invasion;
- entity e is not malicious, open interaction with entity e;
- VWH (the vulnerable world hypothesis), global police state.

"VWH: If technological development continues then a set of capabilities will at some point be attained that make the devastation of civilization extremely likely, unless civilization sufficiently exits the semi-anarchic default condition."[4]

The "the semi-anarchic default condition":

1. limited capacity for preventive policing;
2. limited capacity for global governance;
3. diverse motivations: "There is a wide and recognizably human distribution of motives represented by a large population of actors (at both the individual and state level) - in particular, there are many actors motivated, to a substantial degree, by perceived self-interest (e.g. money, power, status, comfort and convenience) and there are some actors ('the apocalyptic residual') who would act in ways that destroy civilization even at high cost to themselves."[5]

References

Bostrom, N. (2011). Information Hazards: A Typology of Potential Harms from Knowledge. *Review of Contemporary Philosophy*, *10*, 44-79. Citations are from Bostrom's website copy:
https://www.nickbostrom.com/information-hazards.pdf Retrieved 9th Sep. 2020.

Bostrom, N. (2019). The Vulnerable World Hypothesis. *Global Policy*, *10*(4), 455-476. Nov. 2019. Citations are from Bostrom's website copy:
https://nickbostrom.com/papers/vulnerable.pdf Retrieved 24th Mar. 2020.

Brockman, J. (Ed.) (2019). *Possible Minds: Twenty-Five Ways of Looking at AI*. Penguin. ISBN: 978-0525557999. Searches:
https://www.amazon.com/s?k=978-0525557999
https://www.google.com/search?q=isbn+978-0525557999
https://lccn.loc.gov/2018032888

[4] Bostrom (2019) p. 457.
[5] Bostrom (2019) pp. 457-458.

Butler, S. (1863). Darwin among the machines. *The Press (Canterbury, New Zealand)*. Reprinted in Butler et al. (1923).

Butler, S., Jones, H., & Bartholomew, A. (1923). *The Shrewsbury Edition of the Works of Samuel Butler Vol. 1*. J. Cape. No ISBN. https://books.google.com/books?id=B-LQAAAAMAAJ Retrieved 27th Oct. 2020.

Dawkins, R. (2016). *The Selfish Gene*. Oxford, 40th anniv. ed. ISBN: 978-0198788607. Searches: https://www.amazon.com/s?k=9780198788607 https://www.google.com/search?q=isbn+9780198788607 https://lccn.loc.gov/2016933210

Dyson, G. (2020). *Analogia: The Emergence of Technology Beyond Programmable Control*. Farrar, Straus and Giroux. ISBN: 978-0374104863. Searches: https://www.amazon.com/s?k=9780374104863 https://www.google.com/search?q=isbn+9780374104863 https://catalog.loc.gov/vwebv/search?searchArg=9780374104863

Dyson, G. B. (1997). *Darwin Among The Machines: The Evolution Of Global Intelligence*. Basic Books. ISBN: 978-0465031627. Searches: https://www.amazon.com/s?k=978-0465031627 https://www.google.com/search?q=isbn+978-0465031627 https://lccn.loc.gov/2012943208

Kaczynski, T. J., & Skrbina, D. (2010). *Technological Slavery: The Collected Writings of Theodore J. Kaczynski*. Feral House. No ISBN. https://archive.org/details/TechnologicalSlaveryTheCollectedWritingsOfTheodoreJ.KaczynskiA.l TheUnabomber/page/n91/mode/2up Retrieved 11 Jan. 2023.

Koch, C. G. (2007). *The Science of Success*. Wiley. ISBN: 978-0470139882. Searches: https://www.amazon.com/s?k=9780470139882 https://www.google.com/search?q=isbn+9780470139882 https://lccn.loc.gov/2007295977

Kurzweil, R. (1999). *The Age of Spiritual Machines: When Computers Exceed Human Intelligence*. Penguin Books. ISBN: 0140282025. Searches: https://www.amazon.com/s?k=0140282025 https://www.google.com/search?q=isbn+0140282025 https://lccn.loc.gov/98038804

Retraice (2022/10/23). Re27: Now That's a World Model - WM4. *retraice.com*. https://www.retraice.com/segments/re27 Retrieved 24th Oct. 2022.

Retraice (2022/10/27). Re31: What's Happening That Matters - WM5. *retraice.com*. https://www.retraice.com/segments/re31 Retrieved 28th Oct. 2022.

Retraice (2022/11/27). Re63: Seventeen Reasons to Learn AI. *retraice.com*. https://www.retraice.com/segments/re63 Retrieved Monday Nov. 2022.

Russell, B. (1952). *The Impact Of Science On Society*. George Allen and Unwin Ltd. No ISBN. https://archive.org/details/impactofscienceo0000unse_t0h6 Retrieved 15th, Nov.

2022. Searches: https://www.amazon.com/s?k=The+Impact+Of+Science+On+Society+Bertran

https://www.google.com/search?q=The+Impact+Of+Science+On+Society+Bertrand+Russell
https://lccn.loc.gov/52014878

Schneier, B. (2003). *Beyond Fear: Thinking Sensibly About Security in an Uncertain World*. Copernicus Books. ISBN: 0387026207. Searches:
https://www.amazon.com/s?k=0387026207
https://www.google.com/search?q=isbn+0387026207
https://lccn.loc.gov/2003051488
Similar edition available at:
https://archive.org/details/beyondfearthinki00schn_0

von Mises, L. (1949). *Human Action: A Treatise on Economics*. Ludwig von Mises Institute, 2010 reprint ed. ISBN: 9781610161459. Searches:
https://www.amazon.com/s?k=9781610161459
https://www.google.com/search?q=isbn+9781610161459
https://lccn.loc.gov/50002445

Wiener, N. (1954). *The Human Use Of Human Beings: Cybernetics and Society*. Da Capo, 2nd ed. ISBN: 978-0306803208. This 1954 ed. missing 'The Voices of Rigidity' chapter of the original 1950 ed. See 1st ed.: https://archive.org/details/humanuse-ofhumanb00wien/page/n11/mode/2up. See also Brockman (2019) p. xviii. Searches for the 2nd ed.:
https://www.amazon.com/s?k=9780306803208
https://www.google.com/search?q=isbn+9780306803208
https://lccn.loc.gov/87037102

Part 2: Visions of Loss

Transcript

Retrace 114 for Thursday, January 12th, 2023 talking about technological dangers Part 2. Yesterday we said that beliefs and the feelings they cause, determine what chances we take, but possibilities don't care about our feelings.

And today we're going to talk about freedom. Say it like that. Freedom. We're going to talk about freedom and losing freedom. What does it look like? I predicted that freedom is going to decrease over the next 20 years and what's gonna happen after that? Nobody does. Nobody even knows next 20. Years, but. What does that decrease going to look like?

Got a few people here. Who are going to try to tell us? Mostly mathematicians and philosophers, one terrorist. All right, let's dive in.

Mathematician Wiener

Norbert Weiner, mathematician. I guess the question we should be asking as we read through this passage is, is this what's at stake in the struggle for freedom of thought and communication? And, by the way, this excerpt is from the censored version of his human use of human beings. So. I cite 54 in the notes, but you can get to the 1950 edition that has voices of rigidity that chapter. It's not in the later editions. They took it out. OK, this is from that chapter I have said before that man's future on Earth will not be long unless man rises to the full level of his inborn powers. For us to be less than a man is to be less. Than alive. Those who are not fully alive do not live long, even in their world of shadows. I have said, moreover, that for a man to be alive is for him to participate in a worldwide scheme of community. Nation it is to have the liberty to test new opinions and to find which of them point somewhere and which of them simply confuse us. It is to have the variability to fit into the world in more places than one. The variability which may lead us to have soldiers when we need soldiers, but which also leads us to have Saints when we need Saints. It is precisely this variability and this communicative integrity of man which I find to be violated and crippled by the present tendency to huddle together according to a comprehensive prearranged plan which is handed to us from above. We must cease to kiss the whip. That lashes us. He continues. There's something in personal holiness which is akin to an act of choice and the word heresy is nothing but the Greek word for choice. Thus your Bishop, however much he may respect the dead, St. can never feel too friendly toward a living one. This brings up a very interesting remark which Professor John von Neumann has made

to. He has said that in modern society, the era of the primitive church is passing. In modern science. In modern science, the era of the primitive church is passing and that the ear of the Bishop is upon us. Indeed, the Heads of great laboratories are very much like bishops with their association with the powerful in all walks of life and the dangers they incur from the Cardinal. Sins of pride and of lust for power, on the other hand, the independent scientist who is worth the slightest consideration as a scientist has a has a consecration which comes entirely from within him, within himself a vocation which demands the possibility of supreme self sacrifice. Price I have indicated that freedom of opinion at the present time is being crushed between the two rigidities of the Church and the Communist Party. Remember, he's writing in the late 1940s. In the United States, we are in the process of developing a new rigidity which combines the methods of both while partaking of the emotional fervor of neither. Our Conservatives of all shades of opinion have somehow got together to make American capitalism and the fifth freedom economic freedom of the businessman supreme throughout all the. World it is this simple attack on our liberties which we must resist if communication is to have the scope that it properly deserves as the central phenomenon of society, and if the human individual is to reach and to maintain his full stature, it is again the American worship of no how as opposed to know what that hampers us. He's writing about communication and control theory before those were really understood as a thing. He was one of the people who brought it to life. So communication is big in his thinking and human use, and then also the book before it. Cybernetics, which is the more mathematical version of it. OK. So that's from the 50s.

Mathematician and philosopher Russell

Another passage from the 50s we've quoted quoted this partially from Bertrand Russell. We'll just give the full quote now we, we quoted it previously. I forget which one I had it in here. Note to Re49. OK, this in the next one were previously quoted in Ref. 49 the question we should ask is will this happen? OK, here's Russ.

It is to be expected that advances. In Physiology and psychology will give governments much more control over individual mentality than they now have. Even in totalitarian countries fit to laid it down, that education should aim to aim at destroying free will, so that after pupils have left school, they should be incapable throughout the rest of their lives of thinking or acting otherwise. Than as their schoolmasters would have. Finished. But in this day this. But in his day, this was an unattainable ideal. What he regarded as the best system in existence produced Karl Marx. In the future, such failures are not likely to occur where there. Dictatorship, diet injections and injunctions will combine from a very early age to produce the character and the beliefs that the authorities consider desirable, and any serious criticism of the powerful of the powers that be will become psychologically impossible, even if all are miserable, all will believe themselves happy. Because the government will tell them that they are

so. And we'll just. Preview here that Ted Kaczynski, our terrorist and mathematician, says similar things throughout his. His manifesto, the thing that got published at his insistence while he was on his terrorist reign. We'll get to that in a second.

Philosopher Horesh

Just quickly. Theo Harash and philosopher, let's ask the question, is this really happening? Today, meanwhile, a previously unimaginable level of thought control is fast being made accessible for every middle income autocracy that chooses to use it. Visit the wrong website and your Social credit score declines, look up the wrong book and it drops further. Mention the wrong phrases on social media and it sinks so low that alarms go off in the camera rooms. When your face flashes on the screen. The opportunities this presents for behavioral modification are simply astonishing as the as the exploration of every forbidden idea or acquaintance can be made part. Of a social credit score. Whose every drop causes another shock in the hearts of the lowly ranked. Yet whether or not China goes so far, they have developed the tools needed to implement a security regime more totalitarian than even that of the East German Stasi at a fraction of the effort and far lower cost for any autocrat who chooses to go for that and go that far, Russians and Turks, Poles and Hungarians could soon. Find themselves entering a vice from which they never escape. For once, such for once such a security regime is implemented, resistance can be shut down in ways not previously imagined. While independent thinking is gradually snuffed out. And he's writing in 2020. I think it's called the fascism, this time that book harsh 2020, the fascism this time. Yes. OK. Really, really good book. OK. Lots of great insights on that.

Mathematician and terrorist Kaczynski

OK, let's listen to Ted Kaczynski, the terrorist, I think. If you can just forget that he killed 3 people and like ruined the lives of 23 others, not to mention the friends and family. We're not listening to him because he's particularly his writing is particularly brilliant. He's actually quoted in Kurzweil spiritual machines and that quote led me to look further into what he said. He's a smart guy, but he also has a lot of the the. He he falls into a lot of the pits, the pitfalls that smart people do. They tend to think they're smarter than they are, and they tend to selectively use history and selectively use logic to arrive at whatever conclusion they were leaning toward anyway, based on like their adolescence. Honestly, I mean, the guy had a a rough life as being a. Super smart kid, and then eventually turned into. A total *******. And but nonetheless, nonetheless his. So his vision stands out and I'm not the only person quoting it, so we're giving it some attention here with those provisos. So we can ask the question, are these the only possible conclusions of industrial society? This is a he was writing this. This was

published in the New York Times and. 9496 that time frame, but it was published later as a book, at least in 2010, and maybe before that. But you can get your hands on it and published. Now, OK, paragraph 172 from his manifesto.

First let us postulate that the computer scientists succeed in developing intelligent machines that can do all things better than human beings can do them. In that case, presumably all work will be done by the by vast, highly organized systems of machines, and no human effort will be necessary. Either of two cases might occur. The machines might be permitted to make all of their own decisions without human oversight, or else human control over the machines might be retained. So. Those are his. Two scenarios. OK, so we either hand it all over to the machines, or some human control remains and he's going to elaborate on what those two things look like. Paragraph 173, if the machines are permitted to make all their own decisions, we can make we can't make any conjecture as to the results because it is impossible to get to guess how such machines might be. Have we only point out that the fate of the human race would be at the mercy of the machines? Sorry to keep interrupting, Mr. Kaczynski, but this, we've gone over this so many times I didn't feel the need to. Stuart Russell. Bostrom, Bostrom site. The people who have given the most articulate voice to this scenario of losing control to the machines. And anyway, that's that's he's talking about it in the 90s. I, Jay good talked about it long between the 60s and von Neumann, maybe in the 40s or 50s. OK, so he's he's not original in this, but it is. He's early and and George Dyson and and the other guy anyway. OK, we only point out that the fate of the human race will be at the mercy of the machines. It might be argued that the human race would never be foolish enough to hand over all power to the machines, but we are suggesting neither that the human race would voluntarily turn power over to the machines, nor that the machines would willfully, would fully seize. Power. What we do suggest is that the human race might easily permit itself to drift into a position of such dependence on the machines that it would have no practical choice. The human race would have no practical choice but to accept all of the machines decisions as society and the other problems that face it become more and more complex. And as machines become more and more intelligent. People will let machines make more and more of their decisions for them simply because machine made decisions will bring better results than man made ones. Eventually, the stage may be reached at which the decisions necessary to keep the systems running. Will the system running will be so complex that human beings will be incapable of making them intelligently at that stage the machines will be in effective control. People won't be able to just turn off the machines because they will be so dependent on them that they that turning them off would amount to suicide and he continues into paragraph. 74. On the other hand, it is possible that human control over the machines may be retained. In that case, the average man may have control over certain private machines of his own, such as his car or his personal computer. But control over large systems of machines will be in the hands of a tiny elite, just as it is today, but with two differences due to improved techniques, the elite will have. Greater control over the masses and because human

work will no longer be necessary, the masses will be superfluous and a useless burden on the system. If the elite is ruthless, they may simply decide to exterminate the mass of. Unity, if they are humane, they may use propaganda or other psychological or biological techniques to reduce the birth rate until the mass of humanity becomes extinct, leaving the world to the elite. Or if the elite consists of soft hearted liberals, they may decide to play the role of good shepherds to the rest of humanity and the human race they will see to it that everyone's physical needs. Are satisfied that all. Children are raised under Psycho. Logically hygienic conditions that everyone has a wholesome hobby to keep them busy, and that anyone who may become dissatisfied under those under dissatisfied undergoes treatment and quotes to cure his problem in quotes. Of course, life will be so purposeless that people will have to be biologically or psychologically engineered either. Remove their need for the the power process. That's a technical term. He uses the power process or to make them sublimate their drive for power into some harmless hobby. These engineered human beings may be happy in such a society, but they most certainly will not be. Free they will. Have they will have been reduced to the status of domestic animals. And that's also quoted in Kurzweil 1999, as I said, OK. I I didn't. I don't want to do a lot of commentary. We'll get to the commentary later in later segments.

Philosopher Bostrom

OK. Finally, Nick Bostrom in his the vulnerable world hypothesis. Sketches this vision. Of how we might deal with the threat of technology. This is a different this is a solution to a different problem. This isn't the machines taking over, this is. This is the risk of machines taking over leads us to basically the same result. So let's have a look. To secure ourselves against civilization. We should ask ourselves to secure ourselves against civilization. Ending new technologies. Would we accept the follow? So this is Bostrom. For a picture of what a really intensive level of surveillance could look like, consider the following vignette. And this is intensive surveillance would be necessary to secure ourselves against the technological threat of unknown apocalyptic technology. And he calls it high tech panopticon. Everybody is fitted with a freedom. Tag A sequent to the more limited wearable surveillance devices familiar today, such as the ankle tag used in several countries as a prison alternative. The body cams worn by many police forces, the pocket trackers and wristbands that some parents use to keep track of their children, and of course, the ubiquitous cell phone, which has been characterized as a personal tracking device that. Can also be used to make. Phone calls the freedom Tag is a slightly more advanced appliance worn around the neck and bedecked with multidirectional cameras and microphones, encrypted video and audio is continuously uploaded from the device to the cloud and machine, interpreted in real time. AI algorithms classify the activities of the wearer. His hand movements nearby objects. In other situational cues, if suspicious activity is detected, the feed is relayed

to one of several patriot monitoring stations. These are vast office complexes staffed 24/7. There are there, are there a Freedom officer reviews, the video footage, video feed on several screens and listens to the audio and headphones. The Freedom officer then determines an appropriate action, such as contacting the tag wearer via an audio link, to ask for explanation or to request a better view. The Freedom officer can also dispatch and inspector a police rapid response unit or a drone to investigate further. In the small fraction of cases where aware or refuses to desist from the prescribed activity after repeated. Things and arrests may be made or other suitable penalties imposed. Citizens are not permitted to remove the freedom tag, except while they are in environments that have been outfitted with adequate external sensors. Which, however, includes most indoor environments and motor vehicles. The system offers fairly sophisticated privacy protections, such as automated blurring of. Intimate body parts and it provides the option to redact identity, revealing data such as faces and name tags, and release it only when the information is needed for an investigation. Both AI enabled mechanisms and human oversight closely monitor all the actions of the freedom officers to prevent abuse. OK, I'm giving a little bit of a wry smile there at the end. That's it for today. These are the visions that we should consider when we talk about technological danger, and we will do so going forward, signing off.

Notes

Human loss of freedom by deference to authority, dependency on machines, and delegation of defense.

Wiener: freedom of thought and opinion, and communication, as vital; Russell: diet, injections and injunctions in the future;
Horesh: technological behavior modification in the present; terrorist Kaczynski: if AI succeeds, we'll have machine control or elite
control, but no freedom; Bostrom: wearable surveillance devices and power in the hands of a very few as solution.

Air date: Thursday, 12th Jan. 2023, 10:00 PM Eastern/US.

All bold emphasis added.

Mathematician Wiener

Is this what's at stake, in the struggle for freedom of thought and communication? Wiener (1954), p. 217:[1]

> "I have said before that **man's future on earth will not be long unless man rises to the full level of his inborn powers.** For us, to be less than a man is to be less than alive. Those who are not fully alive do not live long even in their world of shadows. I have said, moreover, that for man to be alive is for him to participate in a world-wide scheme of **communication**. It is to have the liberty to test new **opinions** and to find which of them point somewhere, and which of them simply confuse us. It is to have the variability to fit into the world in more places than one, the variability which may lead us to have soldiers when we need soldiers, but which also leads us to have saints when we need saints. **It is precisely this variability and this communicative integrity of man which I find to be violated and crippled by the present tendency to huddle together according to a comprehensive prearranged plan,** which is

[1] The following are excerpts from the 1950 edition, within the later-removed chapter *Voices of Rigidity*. See References for a hyperlink.

handed to us from above. We must cease to kiss the whip that lashes us. ..."

p 226: "There is something in personal holiness which is akin to an act of **choice**, and the word *heresy* is nothing but the Greek word for choice. Thus your Bishop, however much he may respect a dead Saint, can never feel too friendly toward a living one.

"This brings up a very interesting remark which Professor John von Neumann has made to me. He has said that in modern science the era of the primitive church is passing, and that the era of the Bishop is upon us. Indeed, the heads of great laboratories are very much like Bishops, with their association with the powerful in all walks of life, and the dangers they incur of the carnal sins of pride and of lust for power. On the other hand, the independent scientist who is worth the slightest consideration as a scientist, has a consecration which comes entirely from within himself: a vocation which demands the possibility of supreme self-sacrifice "

p. 228: "I have indicated that **freedom of opinion** at the present time is being crushed between the two rigidities of the Church and the Communist Party. In the United States we are in the process [1950] of developing a new rigidity which combines the methods of both while partaking of the emotional fervor of neither. Our Conservatives of all shades of opinion have somehow got together **to make American capitalism and the fifth freedom economic freedom[2] of the businessman** supreme throughout all the world "

p. 229: "It is this triple attack on our **liberties** which we must resist, if **communication** is to have the scope that it properly deserves as the central phenomenon of society, and if the human individual is to reach and to maintain his full stature. It is again the American worship of **know-how** as opposed to **know-what** that hampers us."

Mathematician and philosopher Russell

Will this happen?
Russell (1952), pp. 65-66:[3]

"It is to be expected that advances in physiology and psychology will give governments much more control over individual mentality than they now have even in totalitarian countries. Fichte laid it down that education should aim at destroying free-will, so that, after pupils have left school,

[2] https://en.wikipedia.org/wiki/Fifth_Freedom
[3] Previously quoted, in part, in Re49 (Retraice (2022/11/13)).

they shall be incapable, throughout the rest of their lives, of thinking or acting otherwise than as their schoolmasters would have wished. But in his day this was an unattainable ideal: what he regarded as the best system in existence produced Karl Marx. In [the] future such failures are not likely to occur where there is dictatorship. **Diet, injections, and injunctions will combine, from a very early age, to produce the character and the beliefs that the authorities consider desirable, and any serious criticism of the powers that be will become psychologically impossible.** Even if all are miserable, all will believe themselves happy, because the government will tell them that they are so."

Kaczynski says similar things throughout his 'manifesto'.

Philosopher Horesh

Is this really happening already?
Horesh (2020), p. 158:[4]

"Meanwhile, a previously unimaginable level of thought control is fast being made accessible for every middle-income autocracy that chooses to use it. Visit the wrong website and your social credit score declines, look up the wrong book and it drops further, mention the wrong phrases on social media and it sinks so low that alarms go off in the camera rooms when your face flashes on the screen. **The opportunities this presents for behavioral modification are simply astonishing**, as the exploration of every forbidden idea or acquaintance can be made part of a social credit score, whose every drop causes another shock in the hearts of the lowly ranked Yet, whether

or not China goes so far, they have developed the tools needed to implement a security regime more totalitarian than even that of the East German Stasi, at a fraction of the effort and far lower cost, for any autocrat who chooses to go that far. Russians and Turks, Poles and Hungarians, could soon find themselves entering a vise from which they never escape. For once such a security regime is implemented, resistance can be shut down in ways not previously imagined, while independent thinking is gradually snuffed out."

Mathematician and terrorist Kaczynski

Are these the only possible conclusions of industrial society?

[4] Previously quoted in Re49 (Retraice (2022/11/13)).

(Try to forget that Kaczynski killed three people and ruined many more lives. His vision of the future is quoted by many because it is nuanced and sharply observed; it is worth salvaging from the wreckage of his life.)

Kaczynski & Skrbina (2010), pp. 93-94:[5]

> "172. First let us **postulate that the computer scientists succeed** in developing intelligent machines that can do all things better than human beings can do them. In that case presumably all work will be done by vast, highly organized systems of machines and no human effort will be necessary. **Either of two cases might occur.** The machines might be permitted to make all of their own decisions without human oversight, or else human control over the machines might be retained.
>
> "173. If the machines are permitted to make all their own decisions we can't make any conjecture as to the results, because it is impossible to guess how such machines might behave. We only point out that the fate of the human race would be at the mercy of the machines. It might be argued that the human race would never be foolish enough to hand over all power to the machines. **But we are suggesting neither that the human race would voluntarily turn power over to the machines nor that the machines would will fully seize power. What we do suggest is that the human race might easily permit itself to drift into a position of such dependence on the machines that it would have no practical choice but to accept all of the machines' decisions. As society and the problems that face it become more and more complex and as machines become more and more intelligent, people will let machines make more and more of their decisions for them, simply because machine-made decisions will bring better results than man-made ones. Eventually a stage may be reached at which the decisions necessary to keep the system running will be so complex that human beings will be incapable of making them intelligently. At that stage the machines will be in effective control. People won't be able to just turn the machines off, because they will be so dependent on them that turning them off would amount to suicide.**
>
> "174. On the other hand it is possible that human control over the machines may be retained. In that case the average man may have control over certain private machines of his own, such as his car or his personal computer, but **control over large systems of machines will be in the hands of a tiny elite—just as it is today, but with two differences. Due to improved techniques the elite will have greater**

[5] Also quoted in Kurzweil (1999) pp. 179-180.

control over the masses; and because human work will no longer be necessary the masses will be superfluous, a useless burden on the system. If the elite is ruthless they may simply decide to exterminate the mass of humanity. If they are humane they may use propaganda or other psychological or biological techniques to reduce the birth rate until the mass of humanity becomes extinct, leaving the world to the elite. Or, if the elite consist of soft-hearted liberals, they may decide to play the role of good shepherds to the rest of the human race. They will see to it that everyone's physical needs are satisfied, that all children are raised under psychologically hygienic conditions, that everyone has a wholesome hobby to keep him busy, and that anyone who may become dissatisfied undergoes 'treatment' to cure his 'problem.' Of course, life will be so purposeless that people will have to be biologically or psychologically engineered either to remove their need for the power process or to make them 'sublimate' their drive for power into some harmless hobby. These engineered human beings may be happy in such a society, but they most certainly will not be free. They will have been reduced to the status of domestic animals."

Philosopher Bostrom

So far we have heard about losing power and freedom to machines or their controllers. Now we hear about what *preventing* (or trying to prevent) such losses might look like.

To secure ourselves against civilization-ending new technologies, would we accept the following? Would it work?

Bostrom (2019), pp. 465-466:

> "For a picture of what a really intensive level of surveillance could look like, consider the following vignette:
>
> "*High-tech Panopticon*
>
> "Everybody is fitted with a **'freedom tag'**— a sequent to the more limited wearable surveillance devices familiar today, such as the ankle tag used in several countries as a prison alternative, the bodycams worn by many police forces, the pocket trackers and wristbands that some parents use to keep track of their children, and, of course, the ubiquitous cell phone (which has been characterized as 'a personal tracking device that can also be used to make calls'). The freedom tag is a slightly more advanced appliance, **worn around the neck** and bedecked with multidirectional cameras and microphones. Encrypted video and audio is continuously uploaded from the device to the cloud and machine-interpreted in real time. AI algorithms classify the activities of the wearer, his hand movements, nearby objects, and other situational cues. If suspicious activity is detected, the feed is

relayed to one of several **patriot monitoring stations. These are vast office complexes, staffed 24/7**. There, a **freedom officer** reviews the video feed on several screens and listens to the audio in headphones. The freedom officer then determines an appropriate action, such as contacting the tag-wearer via an audiolink to ask for explanations or to request a better view. The freedom officer can also dispatch an inspector, a police rapid response unit, or a drone to investigate further. In the small fraction of cases where the wearer refuses to desist from the proscribed activity after repeated warnings, an arrest may be made or other suitable penalties imposed. Citizens are not permitted to remove the freedom tag, except while they are in **environments that have been outfitted with adequate external sensors (which however includes most indoor environments and motor vehicles)**. The system **offers fairly sophisticated privacy protections**, such as automated blurring of intimate body parts, and it provides the option to redact identity-revealing data such as faces and name tags and release it only when the information is needed for an investigation. **Both AI-enabled mechanisms and human oversight closely monitor all the actions of the freedom officers to prevent abuse.**"

References

Bostrom, N. (2019). The vulnerable world hypothesis. *Global Policy*, *10*(4), 455-476. Nov. 2019. Citations are from Bostrom's website copy: https://nickbostrom.com/papers/vulnerable.pdf Retrieved 24th Mar. 2020.

Brockman, J. (Ed.) (2019). *Possible Minds: Twenty-Five Ways of Looking atAI*. Penguin. ISBN: 978-0525557999. Searches: https://www.amazon.com/s?k=978-0525557999 https://www.google.com/search?q=isbn+978-0525557999 https://lccn.loc.gov/2018032888

Horesh, T. (2020). *The Fascism this Time: and the Global Future of Democracy*. Cosmopolis Press, Kindle ed. ISBN: 0578732939. Searches: https://www.amazon.com/s?k=0578732939 https://www.google.com/search?q=isbn+0578732939

Kaczynski, T. J., & Skrbina, D. (2010). *Technological Slavery: The Collected Writings of Theodore J. Kaczynski*. Feral House. No ISBN. https://archive.org/details/TechnologicalSlaveryTheCollectedWritingsOfTheodoreJ.KaczynskiA.

TheUnabomber/page/n91/mode/2up Retrieved 11 Jan. 2023.

Kurzweil, R. (1999). *The Age of Spiritual Machines: When Computers Exceed Human Intelligence*. Penguin Books. ISBN: 0140282025. Searches:
https://www.amazon.com/s?k=0140282025
https://www.google.com/search?q=isbn+0140282025
https://lccn.loc.gov/98038804

Retraice (2022/11/13). Re49: China is Not F-ing Around. *retraice.com*.
https://www.retraice.com/segments/re49 Retrieved 15th Nov. 2022.

Russell, B. (1952). *The Impact Of Science On Society*. George Allen and Unwin Ltd. No ISBN.
https://archive.org/details/impactofscienceo0000unse_t0h6 Retrieved 15th, Nov. 2022. Searches:
https://www.amazon.com/s?k=The+Impact+Of+Science+On+Society+Bertrand+Russell
https://www.google.com/search?q=The+Impact+Of+Science+On+Society+Bertrand+Russell
https://lccn.loc.gov/52014878

Wiener, N. (1954). *The Human Use Of Human Beings: Cybernetics and Society*. Da Capo, 2nd ed. ISBN: 978-0306803208. This 1954 ed. missing 'The Voices of Rigidity' chapter of the original 1950 ed. See 1st ed.:
https://archive.org/details/humanuseofhumanb00wien/page/n11/mode/2up. See also Brockman (2019) p. xviii. Searches for the 2nd ed.:
https://www.amazon.com/s?k=9780306803208
https://www.google.com/search?q=isbn+9780306803208
https://lccn.loc.gov/87037102

Part 3: Technological Progress, Defined

Transcript

January 13th, 2023, Friday the 13th.

Listen. I gotta say. If I've seen flat in the last few segments. It's not quite I'm not flat. I'm doing fine. I it's it's low power mode. I'm trying to figure this thing out. If I just let myself go, if I really get into it, then I have a ton of extra work because I bring in all these things that I didn't. Plan on talking about and I have a ton of work to write them up in the notes. So low power mode requires. Me to not to be. Flat, but it's the first thing I'm trying. But it's not. It's not that good. It's not as good to. Watch me flat. It looks like I'm unhappy. I'm not. Unhappy. I'm doing great. OK, just had to get that out of the way.

Progress, 'we' and winners

If we're talking about danger, Technological Danger, Part 3. That's what we're. What we're doing here, we're ultimately talking about the. Question of whether or. Not to proceed whether or not to go forward with something technological danger, whether or not to go forward with technology. I we need a definition of technological progress. Just so happens I have one. And this is not going to be the... I'm not a philosopher. I'm not on anything any of the things that I try to do math code, AI, philosophy. Nothing. I'm not. I'm none of those. Things I'm a. Podcaster.

But anyway, look go look for. A good definition of it, technological progress or good versus bad. RTFM is literally the best model of good versus bad that I'm aware of. But that's because I'm not aware of that many. So anyway, technological progress needs a definition. Let's give it one. This is just a draft. Of course it is, but it's like. This enables us to move forward with something that can be improved if we don't write down the definition. OK, shut up. Do it. Don't tell me to shut up.

Let's start with the idea of progress, the word progress and and just remember that what we're getting at here, how, how how would we decide given predictions whether to. Continued technological advance. OK, that's what we're going for here and try to ignore the stuff on the side. I can't anyway.

OK. Progress can't be control over the environment. We can't define progress as control of the environment. Because who's control is? It who is we? We can't all equally control the environment or equally benefit from controlling the environment, and certainly we don't all prefer the same things that's actually. A huge problem. It's a huge, huge problem. You don't. I can't even say the things that that implies because

they'll be taken out of context and may be used to make. Me look bad. Because someday someone will watch these YouTube videos, OK?

This corresponds roughly to Russell Norwich chapter 27. We've talked about a little bit in Re111, problems with the problems with complexity and inconsistency in human preferences, and also boss from Chapter 13, Super Intelligence, Chapter 13 of the problem of locking in forever the prejudices and preconceptions of the present. Generation a possible solution Bostrom mentions, and Yuki goes into detail Bostrom mentions you mentions Lukowski. His idea of coherent extrapolated volition. It's basically the idea of like, what would we tell the machine? This is just tell the machines. This isn't all the other technical, technological progress come danger that we could be talking about. Our progress or? Advanced let's say technological advance come danger that we should be worried about or could be worried about. They're just talking about the machines taking over. And Bukawski in 2004 is saying we should program into the machines, the, the, the search for what we would want if we were smarter, better had lived longer and convert. I haven't read Lukowski 2004, so I can't. I can't go into detail, but I got the gist of it from Bostrom and maybe I'll read it soon. What would we want if we knew ourselves better if we knew the world better? Blah blah, that's what the machine should pursue that. So that's coherent extrapolated volition because we can't put an easy term on the top of that idea. If our volition doesn't converge, which is my way of saying I don't know how kawski deals with it. If if if we don't. If you keep giving human groups. Or or extrapolating what they want or what they would want if they knew more and had lived longer and had more experience. If their wants don't eventually start to align converge over time they diverge or stay, you know Euclidean parallel wherever this I think this entails winners and losers. It has to, right? I mean, it's I forgot to. Put this in. The notes, but it actually makes me think that. Sorry, I'm distracted by something here myself in this window that's delayed and I see my hands. I'm like, what is he doing with that? It it implies that the I got to retrace my steps. Now if it doesn't, it implies winners and losers. Oh, yeah, the the thing that we value about human beings is not it we this is obvious. We say in a certain way we value human beings. More or less based on what they believe, what's in their minds. And what they desire. But beliefs and desires. So we we. Don't value all humans the same, even though we like to think that we could or should or do. If someone believes that, for example, that you should be killed because of your religious beliefs, you value that person less. Unless you really contort your values and say no, no, no, I want this person to have just as much of the world and of life and opportunity that I do. No, no. I mean if the guy wants to. Kill me. **** him. Right. OK. So anyway, that's that's. A little bit on progress and and if we don't converge, if we don't. All end up wanting more or less the same thing, or something that one machine or one human group could pursue over time, given more experience and more understanding than. Then one team's gonna have to win, and one team's gonna have to lose or lots of teams are gonna have to lose. So what team are you on?

Better and worse problems can be empirical

OK. I just want to point out here the better and worse problems can be empirical. So choose between A&B. Carcinogenic bug spray. Think off. I don't know if off is carcinogenic, but I tend to think all that stuff is on some level. At some point. You live long enough and that that's going to get you cancer wise. So carcinogenic bug spray or malaria. Malaria is a short term risk of death, carcinogenic, anything is a long term risk of death. OK, which ones? Which one's a better problem to have? Well, I think it's the the bug spray just because of the time it's not I don't want to get cancer, but I. But I think malaria would kill me quicker. And and maybe more. Will there be more suffering in any way any? Any years of lost life is is worse than than not losing those years. Choose between lead in the water and. Your water supply, I think Flint, MI over the last few years. Between that problem. And fetching pails of water, Jack and Jill style now. Think about this for a. Second, you don't have water in your dwelling. Unless you fetch a pail of it from. The river, the lake, whatever. Or you can have water in your dwelling every. Day but. Every once in a while you're going to get. Screwed and have led. The water and by. Every once in a while, we mean. In Western countries very rarely, but the people who get screwed will not care how rare it is, as they don't in Flint, MI. OK, choose between those two. Finally choose between an unhappy day job. Most of you will be able to identify with this an unhappy day job. And I mean really unhappy. Like, how unhappy have you been with your day job? It's a very relative thing. And OK, that's one problem you might have. Or you can you can reject that problem and accept its alternative. No home utilities. Or no home. So no electricity, no heat. Might change where you decide to live. There are climates where you can live without electricity and heat, but there are certainly many that without air conditioning, heat or air conditioning, but many that you can't. These days, maybe more and more that you can't, given climate change, global warming, so either you. Go work for the man. What's his name? Schaumburg. Schaumburg. What's the guy? 'S name in. The office space. You either go work for Schaumburg. That would be great or no home utilities and maybe no home. Or at least you have to make your own home so you can. This is empirical. You can ask people this question, they can choose right now. You can't ask people in the past and you can't ask people in the future, but you can ask them today. OK, so it's better and worse. Problems can be empirical. So problems and whether they're better or worse can be empirical.

Technological progress

Now let's try and define technological progress. First, let's distinguish distinguish between the ideas of advancing and progressing. So let's call advancing when we just move forward doesn't matter if it's good for us and progressing, which is like we move. Forward and it's. Good for us predict. And so I think in order to define technological

progress, we would say that there are two. Two, we need 2. Clinicians, because we're going to be making decisions, so we need to 1st define predicted progress. Predictive progress is this a technology seems like progress. If the predicted problems, it will create are better to have than the predicted problems it will solve according to the humans alive at the time of prediction. This in part tries to deal with this idea that technology creates more problems than it solves, man, and it's not about more or less it's about better or worse. If technology creates 100 really good problems to have and solves one really bad one, if those hundred are still not as bad as the original one, then technology as progress doesn't matter. How many problems it's created. I think that's a lot of what human history. That's I think that's a lot of what technology has been, it's been progress. On the whole, we have to deal with time. We'll talk about that in a second. So that's predicted progress. That's what happens before you make the decision. After you make the decision, you can define given that. OK, so here's. So here's actual progress, which you can't actually work on or use in making a decision because it's only retrospective actual progress. The technology is progress is actual progress. If given an interval of time, this is crucial. Given an interval of time, the problems it created were better to have than the problems it's solved. According to the humans alive during the. Or maybe we could say according to the humans alive at the time of assessment or something like that, the idea is. That if it turns out that humanity gets extinguished by technology, that. More or less made extinct with nothing redeeming about it. We don't get to be uploaded into machines. And so our consciousness is preserved or anything like that. We're just basically all that's left on the earth is just whatever was carved in stone a million years from now. You could. Say over that. Interval interval between the first-hand acts and the artificial super intelligence that turned us all into paper clips technology was was not there. There was no technological progress because it's not progress to go from the happy state of being a hunter gatherer to being non-existent. OK? But if you change the time interval and say ohh was there technological progress between hand axe and 2023 Friday the 13? 13th Friday the 13th, 2023, you might say. Yeah, there's plenty of technology. Nobody's dying of all these stupid diseases and everyone's being lifted out of poverty. Depending on how you measure it, more or less quickly. But it's GDP, gosh, GDP is fascinating only because I read something today anyway. So the interval. Given you have to provide a time interval to decide this question of. Of actual progress so. A technology is progress if given an interval of time. The problems it created were better to have than the problems it solved. According to the humans alive during the time interval and maybe you could just take. A survey starting right after the hand acts or like a week after the hand acts, just ask all the Cavemen you know why you think things are going like for that group, and then you know every hundred. Years, OK, so. That's actual progress. And then we just add 1. More thing. Prediction progress. IE learning is is possible if we if we use. If we do both. These things we. Predict progress, and then we check on actual progress. We can use actual progress to if we track it and absorb it, we can use it to improve future predicted progress. OK, so the two are yin

and Yang. They go together, they're a learning loop. OK, that's it. A lot of things. Little citations here. Couple of retraces Emma Bostrom. Lukowski. That's it. 3115. Same time tomorrow, signing off.

Notes

How we would decide, given predictions, whether to risk continued technological advance.

Danger, decisions, advancing and progress; control over the environment and 'we'; complex, inconsistent and conflicting human preferences; 'coherent extrapolated volition' (CEV); divergence, winners and losers; the lesser value of humans who disagree; better and worse problems; predicting progress and observing progress; learning from predicting progress.

Air date: Friday, 13th Jan. 2023, 10:00 PM Eastern/US.

Progress, 'we' and winners

If the question is about 'danger', the answer has to be a decision about whether to proceed (advance). But how to think about progress?

Let '**advance**' mean moving forward, whether or not it's good for humanity. Let '**progress**' mean moving forward in a way that's good for humanity, by some definition of good.[1]

Progress can't be control over the environment, because *whose* control? (Who is *we*?) And we can't *all* control equally or benefit equally or prefer the same thing. This corresponds to the Russell & Norvig (2020) chpt. 27 problems of the complexity and inconsistency of human preferences,[2] and Bostrom (2014) chpt 13 problem of "locking in forever the prejudices and preconceptions of the present generation" (p. 256).

A possible solution is Yudkowsky (2004)'s 'coherent extrapolated volition'.[3] If humanity's collective 'volition' doesn't converge, this might entail that there has to be a 'winner' group in the game of humans vs. humans.

This implies the (arguably obvious) conclusion that we humans value other humans more or less depending on the beliefs and desires they hold.

[1] Retraice (2022/10/24).

[2] Cf. Russell & Norvig (2020) p. 34 and Re111 (Retraice (2023/01/09)).

[3] See also Bostrom (2014) p. 259 ff.

Better and worse problems can be empirical

Choose between A and B:

- carcinogenic bug spray, malaria;
- lead in the water sometimes (Flint, MI), fetching pales;
- unhappy day job, no home utilities (or home).

Which do *you* prefer? This is empirical, in that we can ask people. We can't ask people in the past or the future; but we can always ask people in the present to choose between two alternative problems.

Technological progress

First, we need a definition of progress in order to make decisions. Second, we need an answer to the common retort that 'technology creates *more* problems than it solves'. 'More' doesn't matter; what matters is whether the new problems, together, are 'better' than the old problems, together.

We need to define two timeframes of 'progress' because we're going to use the definition to make decisions: one timeframe to classify a technology *before* the decision to build it, and one timeframe to classify it *after* it has been built and has had observable effects. It's the difference between *expected* progress and *observed* progress. Actual, observed progress can only be determined retrospectively.

Predicted progress:

A technology *seems like* progress if:

the predicted problems it will create are better to have than the predicted problems it will solve, according to the humans alive at the time of prediction.[4]

Actual progress:

A technology *is* progress if:

given an interval of time, the problems it created were better to have than the problems it solved,

according to the humans alive during the interval.

(The time element is crucial: a technology will be, by definition, progress if *up to a moment in history* it never caused worse problems than it solved; but once it does cause such problems, it ceases to be progress, by definition.)

Predict*ion* progress (learning):

'Actual progress', if tracked and absorbed, could be used to improve future 'predicted progress'.

[4] The demonstrated preferences of those humans? The CEV of them? This is hard.

References

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford. First published in 2014. Citations are from the pbk. edition, 2016. ISBN: 978-0198739838. Searches:
https://www.amazon.com/s?k=978-0198739838
https://www.google.com/search?q=isbn+978-0198739838
https://lccn.loc.gov/2015956648

Retraice (2022/10/24). Re28: What's Good? RTFM. *retraice.com*.
https://www.retraice.com/segments/re28 Retrieved 25th Oct. 2022.

Retraice (2023/01/09). Re111: AI and the Gorilla Problem. *retraice.com*.
https://www.retraice.com/segments/re111 Retrieved 10th Jan. 2023.

Russell, S., & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach*. Pearson, 4th ed. ISBN: 978-0134610993. Searches:
https://www.amazon.com/s?k=978-0134610993
https://www.google.com/search?q=isbn+978-0134610993
https://lccn.loc.gov/2019047498

Yudkowsky, E. (2004). Coherent extrapolated volition. *Machine Intelligence Research Institute*. 2004.
https://intelligence.org/files/CEV.pdf Retrieved 13th Jan. 2023.

Part 4: When Does the Bad Thing Happen?

Transcript

Saturday, January 14th, 2023.

The chain reaction of questions

We were bold enough. To make a prediction that freedom is going to decrease, we're bold enough to define technological progress. We were bold enough. To talk about visions of loss of freedom. We haven't defined our terms yet. But we should be bold enough. Shouldn't we to talk about when the bad things happen, right? We should be able to do that. Feels like we should be able to do that. Like we're we're saying, we got to make decisions. We're going to have to. Our beliefs are going to affect. Whether or not we go along with this whole, the changes in the amounts of freedom that we have. We make decisions we have to. I mean, it's if they're about the future and then what we want to look back at the past. Time frames matter. We talked about at the tail end of. 1:15. This is 1/16. Did I say 116? When does the bad thing happen? That we're predicting or when did the bad thing happen that we learned from in the past? The technological bad thing? Any bad thing when? When did it start? Like when did it start? When did anything start? Well, it started clearly it started on January 14th. The bad thing, that was the day that the retrace guy said the thing that led to all that bad stuff. Using that power, he had no idea he had that. You had no idea I had have. No, it didn't start on January 14th. You gotta go back further cause I. Was like born in the 80s. So really like if anything came, if it was caused by me, it came from the 80s. Well like I I didn't come from nowhere. Cut off that line of thought, but then? nobody came from nowhere. It started with humans. When was the first human? A question I've been asked more than once by. My dear young lynx. Star Wars reference for for you ******** out there. There was no real first human. It was gradual. When when was the 1st? Creature. The thing is, it's an infinite. Chain of when did it start? When did anything start? What's a thing? OK, enough. What we're dealing with is what philosophers call ontology, and I'm going to solve it for you. No, I'm not. I was just kidding how crazy that is that I said that I'm not going to solve ontology and this is gonna be a short segment because I'm not gonna solve ontology. The. Point is, it's very hard to pick out. A point in time when something starts and when it stops. Which leads you it first leads you to. The question of time, like what the hell is time? And then like, how does causality work? But then it's like if we're talking about time like it's different from space. Ohh well, what's space? So yeah, it's like well it's like the

stuff that's not the objects. Well, what are the objects? So it's like this stuff that's not the space. Do we have free will? Are we in control of? Any of this exactly. All that stuff is a freaking tar pit. And yet...

Ontology and treaties for sharing

The need for ontology. Is going to increase. It's going to the need for a precise ontology is going to increase because we're going to be dealing with technologies that have a more precise ontology every step of the way going forward. So we're going, but we can't do we can't deal with. Can I read you something? I happen to read this today in my studies. Very serendipitous. OK so. Ontology. Well, first, I'll read you the Wikipedia definition of ontology. Just in case you you're not quite with me in metafit metaphysics, which is a branch of philosophy. If if you agree with that the use of that word, ontology is the philosophical study of being as well as related. Concepts such as existence, becoming in reality ontology addresses questions like how entities are grouped into categories. And which of these entities exist on the most? Fundamental level. OK, I think of it. That's to what? Being I hate that like the the Continental people use being in time and being claw being I. Hate that stuff. I hate it. I hate it. I just think of. It as like what is, what is there, what's what is. There in the universe, not just stuff, but also like the space around the stuff and then like. The stuff before it was the stuff that it is now and the stuff it'll be later time is there. Free will stuff.

Is there causality or is?

There OK, so. Ontology is like what is there, what is there? Now, with that definition in mind, I'm going to read you from Emma 4E, page 316. This is in the knowledge representation chapter Chapter 10, and they say this, trust me, this is the this is where we need to arrive. OK, we have we have we need. We need to make a treaty. Not my word. Here we go. We should say up. Front that the enterprise of general ontological engineering has so far had only limited success. None of the top AI applications as listed in chapter one make use of a general ontology. They all use special purpose, knowledge engineering and machine learning, social slash political considerations can make it difficult for competing parties to agree on an ontology, as Tom Gruber in 2004 says. Every ontology is a treaty, a social agreement among people with some common motive in sharing. When competing concerns outweigh the motivation for sharing, there can be no common ontology, and we stopped quoting it. Now they're back to them talking, and when when competing concerns outweigh the motivation for sharing, there could be no common ontology. The smaller the number of stakeholders. Think winners and losers. The smaller the number of stakeholders, the easier it is to create an ontology, and thus it is harder to create a general purpose ontology than a limited purpose. One such as the open biomedical ontology Smith at all 2007.

Prediction: the need for precise ontologies is going to increase.

That is pregnant with significance like that passage is just. Holy crap. Hit me more than once, like a ton of. Bricks, we are. Not going to be able to settle. What time is what space? What matter? What causality? What? Free will are. Not right now. No one. And they've been working their butts off like you could this. They talk about it as well in Chapter 10 and in the bibliography for chapter. 10 a lot of people. Have worked on this. They don't. They haven't cracked it. Let's leave it at that. That so that means when we talk about the bad thing happening in the future, the bad technological thing that we're trying to avoid or the bad thing that happened in. The past, we actually don't have. The philosophical foundations that we need to even have that conversation, but we still have to make decisions. We still have to make descriptions and judgments and do all the things that it means to be human. We can't just surrender to a philosophical problem, just cause we haven't solved it yet. But man treaties. And the fewer the number of stakeholders. Gosh, it just makes me think about the game winners and losers. What? What, what? Why do you make a treat? Why do you make an ontological treaty? You have to. Have some common motive in sharing some common motive in sharing. And then that's that's quoting Gruber and then back to the aim of guys when competing concerns outweigh the motivation for sharing. There can be no common ontology. They just declare that. Now I'm I'm I'll defer to them, but the smaller the number of stakeholders, the easier it is to create. An ontology. Think about that. The computer control people. The insiders by some definition. The smaller the number of them who cooperate. The easier it is. To make an ontological treaty and what does an ontological treaty? It enables AI, it enables knowledge representation. It enables work to be done, work to be done. If you can't. Agree on what things there are or what categories. Or blah blah blah. You're going to be sitting in the philosophy department doing squat. While the AI people are in the AI department. About to make you. Go squat. I don't know what that means. Make you go splat. That's it. OK. The only point I really wanted to make is that it's hard. Deciding in ontology. And we're not going to do it. But you and me, we got a good thing going on here. We'll make a treaty every time we need to make a treaty. We'll make a treaty. That's it. 316 signing off.

Notes

Agreements about reality in technological progress.

Basic questions; a chain reaction of philosophy; deciding what is and isn't in the world; agreeing with others in order to achieve sharing; other concerns compete with sharing and prevent agreement; the need for agreement increasing.

Air date: Saturday, 14th Jan. 2023, 10:00 PM Eastern/US.

The chain reaction of questions

We were bold enough to predict a decrease in freedom (without defining it);[1] we were bold enough to define technological progress (*with* defining it).[2] But in predicting and assessing 'bad things' (i.e. technological danger), we *should* be able to talk about *when* the bad things might or might not happen, did or didn't happen. But can we? When does anything start and stop? How to draw the lines in chronology? How to draw the lines in causality? There is a chain reaction of questions and subjects:

- **Time:** *When* did it start? With the act, or the person, or the species?
- **Space:** *Where* did it start?
- **Matter:** What *is* it?
- **Causality:** What *caused* it?
- **Free will:** Do we *cause* anything, really?

Ontology and treaties for sharing

Ontology is the subset of philosophy that deals with 'being', 'existence', 'reality', the categories of such things, etc. I.e., it's about 'what is', or 'What *is* there?', or 'the stuff' of the world. From AIMA4e (emphasis added):

[1] Retraice (2023/01/11)
[2] Retraice (2023/01/13)

> "We should say up front that the enterprise of general ontological engineering has so far had only limited success. None of the top AI applications (as listed in Chapter 1) make use of a general ontology—they all use special-purpose knowledge engineering and machine learning. Social/political considerations can make it difficult for competing parties to agree on an ontology. As Tom Gruber (2004) says, **'Every ontology is a treaty—a social agreement—among people with some common motive in sharing.' When competing concerns outweigh the motivation for sharing, there can be no common ontology. The smaller the number of stakeholders, the easier it is to create an ontology**, and thus it is harder to create a generalpurpose ontology than a limited-purpose one, such as the Open Biomedical Ontology."[3]

Prediction: the need for precise ontologies is going to increase.

Ontology is not a solved problem—neither in philosophy nor artificial intelligence. Yet we can't sit around and wait. The computer control game is on. We have to act and act effectively. And further, *our* need for precise ontologies—that is, the making of treaties—is going to *increase* because we're going to be dealing with *technologies* that have more and more precise ontologies. So, consider:

- More stakeholders makes treaties less likely;
- The problems that we can solve without AI (and its ontologies and our own ontologies) are decreasing;
- Precise ontology enables knowledge representation (outside of machine-learning), and therefore AI, and therefore the effective building of technologies and taking of actions, and therefore work to be done;
- Treaties can make winners and losers in the computer control game;
- Competing concerns can outweigh the motive for sharing, and therefore treaties, and therefore winning.

[3] Russell & Norvig (2020) p. 316. And Gruber's *Every Ontology Is a Treaty* (2004): https://tomgruber.org/writing/sigsemis-2004

References

Retraice (2023/01/11). Re113: Uncertainty, Fear and Consent (Technological Danger, Part 1). *retraice.com*.
https://www.retraice.com/segments/re113 Retrieved 12th Jan. 2023.

Retraice (2023/01/13). Re115: Technological Progress, Defined (Technological Danger, Part 3). *retraice.com*.
https://www.retraice.com/segments/re115 Retrieved 14th Jan. 2023.

Russell, S., & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach*. Pearson, 4th ed. ISBN: 978-0134610993. Searches:
https://www.amazon.com/s?k=978-0134610993
https://www.google.com/search?q=isbn+978-0134610993
https://lccn.loc.gov/2019047498

Retraice
Technological Danger
Jan 12, 2023

<retraice.com>