

Best of LessWrong: November 2018

1. [Act of Charity](#)
2. [Is Clickbait Destroying Our General Intelligence?](#)
3. [Is Science Slowing Down?](#)
4. [Incorrect hypotheses point to correct observations](#)
5. [Embedded Agency \(full-text version\)](#)
6. [Counterintuitive Comparative Advantage](#)
7. [Robust Delegation](#)
8. [Subsystem Alignment](#)
9. [Sam Harris and the Is-Ought Gap](#)
10. [Preschool: Much Less Than You Wanted To Know](#)
11. [Combat vs Nurture: Cultural Genesis](#)
12. [Rationality Is Not Systematized Winning](#)
13. [How rapidly are GPUs improving in price performance?](#)
14. [Hyperreal Brouwer](#)
15. [Acknowledging Human Preference Types to Support Value Learning](#)
16. [Speculations on improving debating](#)
17. [Embedded World-Models](#)
18. [The Inspection Paradox is Everywhere](#)
19. [Specification gaming examples in AI](#)
20. [Clickbait might not be destroying our general Intelligence](#)
21. [Alignment Newsletter #34](#)
22. [What are Universal Inductors, Again?](#)
23. [Update the best textbooks on every subject list](#)
24. [Burnout: What it is and how to Treat it.](#)
25. [How democracy ends: a review and reevaluation](#)
26. [Embedded Curiosities](#)
27. [The Ubiquitous Converse Lawvere Problem](#)
28. [Implementations of immortality](#)
29. [On MIRI's new research directions](#)
30. [Double-Dipping in Dunning-Kruger](#)
31. [Is Copenhagen or Many Worlds true? An experiment. What? Yes.](#)
32. [New safety research agenda: scalable agent alignment via reward modeling](#)
33. [If You Want to Win, Stop Conceding](#)
34. [On Rigorous Error Handling](#)
35. [Clarifying "AI Alignment"](#)
36. [The Vulnerable World Hypothesis \(by Bostrom\)](#)
37. [Stabilize-Reflect-Execute](#)
38. [Topological Fixed Point Exercises](#)
39. [Fixed Point Discussion](#)
40. [Status model](#)
41. [Speculative Evopsych, Ep. 1](#)
42. [Believing others' priors](#)
43. [When does rationality-as-search have nontrivial implications?](#)
44. [Aligned AI, The Scientist](#)
45. [Fixed Point Exercises](#)
46. [Four factors that moderate the intensity of emotions](#)
47. [October gwern.net links](#)
48. [Diagonalization Fixed Point Exercises](#)
49. [Alignment Newsletter #33](#)
50. [Humans can be assigned any values whatsoever...](#)

Best of LessWrong: November 2018

1. [Act of Charity](#)
2. [Is Clickbait Destroying Our General Intelligence?](#)
3. [Is Science Slowing Down?](#)
4. [Incorrect hypotheses point to correct observations](#)
5. [Embedded Agency \(full-text version\)](#)
6. [Counterintuitive Comparative Advantage](#)
7. [Robust Delegation](#)
8. [Subsystem Alignment](#)
9. [Sam Harris and the Is-Ought Gap](#)
10. [Preschool: Much Less Than You Wanted To Know](#)
11. [Combat vs Nurture: Cultural Genesis](#)
12. [Rationality Is Not Systematized Winning](#)
13. [How rapidly are GPUs improving in price performance?](#)
14. [Hyperreal Brouwer](#)
15. [Acknowledging Human Preference Types to Support Value Learning](#)
16. [Speculations on improving debating](#)
17. [Embedded World-Models](#)
18. [The Inspection Paradox is Everywhere](#)
19. [Specification gaming examples in AI](#)
20. [Clickbait might not be destroying our general Intelligence](#)
21. [Alignment Newsletter #34](#)
22. [What are Universal Inductors, Again?](#)
23. [Update the best textbooks on every subject list](#)
24. [Burnout: What it is and how to Treat it.](#)
25. [How democracy ends: a review and reevaluation](#)
26. [Embedded Curiosities](#)
27. [The Ubiquitous Converse Lawvere Problem](#)
28. [Implementations of immortality](#)
29. [On MIRI's new research directions](#)
30. [Double-Dipping in Dunning-Kruger](#)
31. [Is Copenhagen or Many Worlds true? An experiment. What? Yes.](#)
32. [New safety research agenda: scalable agent alignment via reward modeling](#)
33. [If You Want to Win, Stop Conceding](#)
34. [On Rigorous Error Handling](#)
35. [Clarifying "AI Alignment"](#)
36. [The Vulnerable World Hypothesis \(by Bostrom\)](#)
37. [Stabilize-Reflect-Execute](#)
38. [Topological Fixed Point Exercises](#)
39. [Fixed Point Discussion](#)
40. [Status model](#)
41. [Speculative Evopsych, Ep. 1](#)
42. [Believing others' priors](#)
43. [When does rationality-as-search have nontrivial implications?](#)
44. [Aligned AI, The Scientist](#)
45. [Fixed Point Exercises](#)
46. [Four factors that moderate the intensity of emotions](#)
47. [October gwern.net links](#)
48. [Diagonalization Fixed Point Exercises](#)
49. [Alignment Newsletter #33](#)

50. Humans can be assigned any values whatsoever...

Act of Charity

(Cross-posted from [my blog](#))

The stories and information posted here are artistic works of fiction and falsehood.
Only a fool would take anything posted here as fact.

—Anonymous

Act I.

Carl walked through the downtown. He came across a charity stall. The charity worker at the stall called out, "Food for the Africans. Helps with local autonomy and environmental sustainability. Have a heart and help them out." Carl glanced at the stall's poster. Along with pictures of emaciated children, it displayed infographics about how global warming would cause problems for African communities' food production, and numbers about how easy it is to help out with money. But something caught Carl's eye. In the top left, in bold font, the poster read, "IT IS ALL AN ACT. ASK FOR DETAILS."

Carl: "It's all an act, huh? What do you mean?"

Worker: "All of it. This charity stall. The information on the poster. The charity itself. All the other charities like us. The whole Western idea of charity, really."

Carl: "Care to clarify?"

Worker: "Sure. This poster contains some correct information. But a lot of it is presented in a misleading fashion, and a lot of it is just lies. We designed the poster this way because it fits with people's idea of a good charity they should give money to. It's a prop in the act."

Carl: "Wait, the stuff about global warming and food production is a lie?"

Worker: "No, that part is actually true. But in context we're presenting it as some kind of imminent crisis that requires an immediate infusion of resources, when really it's a very long-term problem that will require gradual adjustment of agricultural techniques, locations, and policies."

Carl: "Okay, that doesn't actually sound like more of a lie than most charities tell."

Worker: "Exactly! It's all an act."

Carl: "So why don't you tell the truth anyway?"

Worker: "Like I said before, we're trying to fit with people's idea of what a charity they should give money to looks like. More to the point, we want them to feel *compelled* to give us money. And they are compelled by some acts, but not by others. The idea of an immediate food crisis creates more moral and social pressure towards immediate action, than the idea that there will be long-term agricultural problems that require adjustments."

Carl: "That sounds...kind of scammy?"

Worker: "Yes, you're starting to get it! The act is about violence! It's all violence!"

Carl: "Now hold on, that seems like a false equivalence. Even if they were scammed by you, they still gave you money of their own free will."

Worker: "Most people, at some level, know we're lying to them. Their eyes glaze over 'IT IS ALL AN ACT' as if it were just a regulatory requirement to put this on charity posters. So why would they give money to a charity that lies to them? Why do you think?"

Carl: "I'm not nearly as sure as you that they know this! Anyway, even if they know at some level it's a lie, that doesn't mean they *consciously* know, so to their conscious mind it seems like being completely heartless."

Worker: "Exactly, it's emotional blackmail. I even say 'Have a heart and help them out'. So if they don't give us money, there's a really convenient story that says they're heartless, and a lot of them will even start thinking about themselves that way. Having that story told about them opens them up to violence."

Carl: "How?"

Worker: "Remember Martin Shkreli?"

Carl: "Yeah, that asshole who jacked up the Daraprim prices."

Worker: "Right. He ended up going to prison. Nominally, it was for securities fraud. But it's not actually clear that whatever security fraud he did was worse than what others in his industry were doing. Rather, it seems likely that he was especially targeted because he was a heartless asshole."

Carl: "But he still broke the law!"

Worker: "How long would you be in jail if you got punished for every time you had broken the law?"

Carl: "Well, I've done a few different types of illegal drugs, so... a lot of years."

Worker: "Exactly. Almost everyone is breaking the law. So it's really, really easy for the law to be enforced selectively, to punish just about anyone. And the people who get punished the most are those who are villains in the act."

Carl: "Hold on. I don't think someone would actually get sent to prison because they didn't give you money."

Worker: "Yeah, that's pretty unlikely. But things like it will happen. People are more likely to give if they're walking with other people. I infer that they believe they will be abandoned if they do not give."

Carl: "That's a far cry from violence."

Worker: "Think about the context. When you were a baby, you relied on your parents to provide for you, and abandonment by them would have meant certain death. In the environment of evolutionary adaptation, being abandoned by your band would have been close to a death sentence. This isn't true in the modern world, but people's

brains mostly don't really distinguish abandonment from violence, and we exploit that."

Carl: "That makes some sense. I still object to calling it violence, if only because we need a consistent definition of 'violence' to coordinate, well, violence against those that are violent. Anyway, I get that this poster is an act, and the things you say to people walking down the street are an act, but what about the charity itself? Do you actually do the things you say you do?"

Worker: "Well, kind of. We actually do give these people cows and stuff, like the poster says. But that isn't our main focus, and the main reason we do it is, again, because of the act."

Carl: "Because of the act? Don't you care about these people?"

Worker: "Kind of. I mean, I do care about them, but I care about myself and my friends more; that's just how humans work. And if it doesn't cost me much, I will help them. But I won't help them if it puts our charity in a significantly worse position."

Carl: "So you're the heartless one."

Worker: "Yes, and so is everyone else. Because the standard you're set for 'not heartless' is not one that any human actually achieves. They just deceive themselves about how much they care about random strangers; the part of their brain that inserts these self-deceptions into their conscious narratives is definitely not especially altruistic!"

Carl: "According to your own poster, there's going to be famine, though! Is the famine all an act to you?"

Worker: "No! Famine isn't an act, but most of our activities in relation to it are. We give people cows because that's one of the standard things charities like ours are supposed to do, and it looks like we're giving these people local autonomy and stuff."

Carl: "*Looks like?* So this is all just optics?"

Worker: "Yes! Exactly!"

Carl: "I'm actually really angry right now. You are a terrible person, and your charity is terrible, and you should die in a fire."

Worker: "Hey, let's actually think through this ethical question together. There's a charity pretty similar to ours that's set up a stall a couple blocks from here. Have you seen it?"

Carl: "Yes. They do something with water filtering in Africa."

Worker: "Well, do you think their poster is more or less accurate than ours?"

Carl: "Well, I know yours is a lie, so..."

Worker: "Hold on. This is Gell-Mann amnesia. You know ours is a lie *because I told you*. This should adjust your model of how charities work in general."

Carl: "Well, it's still *plausible* that they are effective, so I can't condemn—"

Worker: "Stop. In talking of plausibility rather than probability, you are uncritically participating in the act. You are taking symbols at face value, unless there is clear disproof of them. So you will act like you believe any claim that's 'plausible', in other words one that can't be disproven from within the act. You have never, at any point, checked whether either charity is doing anything in the actual, material world."

Carl: "...I suppose so. What's your point, anyway?"

Worker: "You're shooting the messenger. All or nearly all of these charities are scams. Believe me, we've spent time visiting these other organizations, and they're universally fraudulent, they just have less self-awareness about it. You're only morally outraged at the ones that don't hide it. So your moral outrage optimizes against your own information. By being morally outraged at us, you are asking to be lied to."

Carl: "Way to blame the victim. You're the one lying."

Worker: "We're part of the same ecosystem. By rewarding a behavior, you cause more of it. By punishing it, you cause less of it. You reward lies that have plausible deniability and punish truth, when that truth is told by sinners. You're actively encouraging more of the thing that is destroying your own information!"

Carl: "It still seems pretty strange to think that they're all scams. Like, some of my classmates from college went into the charity sector. And giving cows to people who have food problems actually seems pretty reasonable."

Worker: "It's well known by development economists that aid generally creates dependence, that in giving cows to people we disrupt their local economy's cow market, reducing the incentive to raise cattle. And in theory it could still be worth it, but our preliminary calculations indicate that it probably isn't."

Carl: "Hold on. You *actually ran the calculation*, found that your intervention was net harmful, and then *kept doing it?*"

Worker: "Yes. Again, it is all—"

Carl: "What the fuck, seriously? You're a terrible person."

Worker: "Do you think any charity *other than us* would have run the calculation we did, and then actually believe the result? Or would they have fudged the numbers here and there, and when even a calculation with fudged numbers indicated that the intervention was ineffective, come up with a reason to discredit this calculation and replace it with a different one that got the result they wanted?"

Carl: "Maybe a few... but I see your point. But there's a big difference between acting immorally because you deceived yourself, and acting immorally with a clear picture of what you're doing."

Worker: "Yes, the second one is much less bad!"

Carl: "What?"

Worker: "All else being equal, it's better to have clearer beliefs than muddier ones, right?"

Carl: "Yes. But in this case, it's *very clear* that the person with the clear picture is acting immorally, while the self-deceiver, uhh.."

Worker: "...has plausible deniability. Their stories are plausible even though they are false, so they have more privilege within the act. They gain privilege by muddying the waters, or in other words, destroying information."

Carl: "Wait, are you saying self-deception is a choice?"

Worker: "Yes! It's called 'motivated cognition' for a reason. Your brain runs something like a utility-maximization algorithm to tell when and how you should deceive yourself. It's epistemically correct to take the intentional stance towards this process."

Carl: "But I don't have any control over this process!"

Worker: "Not consciously, no. But you can notice the situation you're in, think about what pressures there are on you to self-deceive, and think about modifying your situation to reduce these pressures. And you can do this to other people, too."

Carl: "Are you saying everyone is morally obligated to do this?"

Worker: "No, but it might be in your interest, since it increases your capabilities."

Carl: "Why don't you just run a more effective charity, and advertise on that? Then you can outcompete the other charities."

Worker: "That's not fashionable anymore. The 'effectiveness' branding has been tried before; donors are tired of it by now. Perhaps this is partially because there aren't functional systems that actually check which organizations are effective and which aren't, so scam charities brand themselves as effective end up outcompeting the actually effective ones. And there are organizations claiming to evaluate charities' effectiveness, but they've largely also become scams by now, for exactly the same reasons. The fashionable branding now is environmentalism."

Carl: "This is completely disgusting. Fashion doesn't help people. Your entire sector is morally depraved."

Worker: "You are entirely correct to be disgusted. This moral depravity is a result of dysfunctional institutions. You can see it outside charity too; schools are authoritarian prisons that don't even help students learn, courts put people in cages for not spending enough on a lawyer, the US military blows up civilians unnecessarily, and so on. But you already knew all that, and ranting about these things is itself a trope. It is difficult to talk about how broken the systems are without this talking itself being interpreted as merely a cynical act. That's how deep this goes. Please actually update on this rather than having your eyes glaze over!"

Carl: "How do you even deal with this?"

Worker: "It's already the reality you've lived in your whole life. The only adjustment is to realize it, and be able to talk about it, without this destroying your ability to participate in the act when it's necessary to do so. Maybe functional information-processing institutions will be built someday, but we are stuck with this situation for now, and we'll have no hope of building functional institutions if we don't understand our current situation."

Carl: "You are wasting so much potential! With your ability to see social reality, you could be doing all kinds of things! If everyone who were as insightful as you were as pathetically lazy as you, there would be no way out of this mess!"

Worker: "Yeah, you're right about that, and I might do something more ambitious someday, but I don't really want to right now. So here I am. Anyway... food for the Africans. Helps with local autonomy and environmental sustainability. Have a heart and help them out."

Carl sighed, fished a 10 dollar bill from his wallet, and gave it to the charity worker.

Is Clickbait Destroying Our General Intelligence?

(Cross-posted from Facebook.)

Now and then people have asked me if I think that other people should also avoid high school or college if they want to develop new ideas. This always felt to me like a wrong way to look at the question, but I didn't know a right one.

Recently I thought of a scary new viewpoint on that subject.

This started with a conversation with Arthur where he mentioned an idea by Yoshua Bengio about the software for [general intelligence](#) having been developed memetically. I remarked that I didn't think duplicating this culturally transmitted software would be a significant part of the problem for AGI development. (Roughly: low-fidelity software tends to be algorithmically shallow. Further discussion moved to comment below.)

But this conversation did get me thinking about the topic of culturally transmitted software that contributes to human general intelligence. That software can be an *important* gear even if it's an algorithmically shallow part of the overall machinery. Removing a few simple gears that are 2% of a machine's mass can reduce the machine's performance by way more than 2%. Feral children would be the case in point.

A scary question is whether it's possible to do *subtler* damage to the culturally transmitted software of general intelligence.

I've had the sense before that the Internet is turning our society stupider and meaner. My primary hypothesis is "The Internet is selecting harder on a larger population of ideas, and sanity falls off the selective frontier once you select hard enough."

To review, there's a general idea that strong (social) selection on a characteristic imperfectly correlated with some other metric of goodness can be bad for that metric, where weak (social) selection on that characteristic was good. If you press scientists a *little* for publishable work, they might do science that's of greater interest to others. If you select very harshly on publication records, the academics spend all their time worrying about publishing and real science falls by the wayside.

On my feed yesterday was an essay complaining about how the intense competition to get into Harvard is producing a monoculture of students who've lined up every single standard accomplishment and how these students don't know anything else they want to do with their lives. Gentle, soft competition on a few accomplishments might select genuinely stronger students; hypercompetition for the appearance of strength produces weakness, or just emptiness.

A hypothesis I find plausible is that the Internet, and maybe television before it, selected much more harshly from a much wider field of memes; and also allowed tailoring content more narrowly to narrower audiences. The Internet is making it possible for ideas that are optimized to appeal hedonically-virally within a filter bubble to outcompete ideas that have been even slightly optimized for anything else. We're

looking at a collapse of reference to expertise because deferring to expertise costs a couple of hedons compared to being told that all your intuitions are perfectly right, and at the harsh selective frontier there's no room for that. We're looking at a collapse of interaction between bubbles because there used to be just a few newspapers serving all the bubbles; and now that the bubbles have separated there's little incentive to show people how to be fair in their judgment of ideas for other bubbles, it's not *the most* appealing Tumblr content. Print magazines in the 1950s were hardly perfect, but they could get away with sometimes presenting complicated issues as complicated, because there weren't a hundred blogs saying otherwise and stealing their clicks. Or at least, that's the hypothesis.

It seems plausible to me that *basic* software for intelligent functioning is being damaged by this hypercompetition. Especially in a social context, but maybe even outside it; that kind of thing tends to slop over. When someone politely presents themselves with a careful argument, does your cultural software tell you that you're supposed to listen and make a careful response, or make fun of the other person and then laugh about how they're upset? What about when your own brain tries to generate a careful argument? Does your cultural milieu give you any examples of people showing how to really care deeply about something (i.e. debate consequences of paths and how hard to the best one), or is everything you see just people competing to be loud in their identification? The Occupy movement not having any demands or agenda could represent mild damage to a gear of human general intelligence that was culturally transmitted and that enabled processing of a certain kind of goal-directed behavior. And I'm not sure to what extent that is merely a metaphor, versus it being simple fact if we could look at the true software laid out. If you look at how some bubbles are talking and thinking now, "intellectually feral children" doesn't seem like entirely inappropriate language.

Shortly after that conversation with Arthur, it occurred to me that I was pretty much raised and socialized by my parents' collection of science fiction.

My parents' collection of *old* science fiction.

Isaac Asimov. H. Beam Piper. A. E. van Vogt. Early Heinlein, because my parents didn't want me reading the later books.

And when I did try reading science fiction from later days, a lot of it struck me as... icky. *Neuromancer*, bleah, what is *wrong* with this book, it feels *damaged*, why do people like this, it feels like there's way too much flash and it ate the substance, it's showing off way too hard.

And now that I think about it, I feel like a lot of my writing on rationality would be a lot more popular if I could go back in time to the 1960s and present it there. "Twelve Virtues of Rationality" is what people could've been reading instead of Heinlein's *Stranger in a Strange Land*, to take a different path from the branching point that found *Stranger in a Strange Land* appealing.

I didn't stick to merely the culture I was raised in, because that wasn't what that culture said to do. The characters I read didn't keep to the way *they* were raised. They were constantly being challenged with new ideas and often modified or partially rejected those ideas in the course of absorbing them. If you were immersed in an alien civilization that had some good ideas, you were supposed to consider it open-mindedly and then steal only the good parts. Which... kind of sounds axiomatic to me? You could make a case that this is an obvious guideline for how to do *generic*

optimization. It's just what you do to process an input. And yet "when you encounter a different way of thinking, judge it open-mindedly and then steal only the good parts" is directly contradicted by some modern software that seems to be memetically hypercompetitive. It probably sounds a bit alien or weird to some people reading this, at least as something that you'd say out loud. Software contributing to generic optimization has been damaged.

Later the Internet came along and exposed me to some modern developments, some of which are indeed improvements. But only after I had a cognitive and ethical foundation that could judge which changes were progress versus damage. More importantly, a cognitive foundation that had the idea of even *trying* to do that. Tversky and Kahneman didn't exist in the 1950s, but when I was exposed to this new cognitive biases literature, I reacted like an Isaac Asimov character trying to integrate it into their existing ideas about psychohistory, instead of a William Gibson character wondering how it would look on a black and chrome T-Shirt. If that reference still means anything to anyone.

I suspect some culturally transmitted parts of the general intelligence software got damaged by radio, television, and the Internet, with a key causal step being an increased hypercompetition of ideas compared to earlier years. I suspect this independently of any other hypotheses about my origin story. It feels to me like the historical case for this thesis ought to be visible by mere observation to anyone who watched the quality of online discussion degrade from 2002 to 2017.

But if you consider me to be more than usually intellectually productive for an average Ashkenazic genius in the modern generation, then in this connection it's an interesting and scary further observation that I was initially socialized by books written before the Great Stagnation. Or by books written by authors from only a single generation later, who read a lot of old books themselves and didn't watch much television.

That hypothesis doesn't feel wrong to me the way that "oh you just need to not go to college" feels wrong to me.

Is Science Slowing Down?

[This post was up a few weeks ago before getting taken down for complicated reasons. They have been sorted out and I'm trying again.]

Is scientific progress slowing down? I recently got a chance to attend a conference on this topic, centered around a paper by [Bloom, Jones, Reenen & Webb \(2018\)](#).

BJRW identify areas where technological progress is easy to measure – for example, the number of transistors on a chip. They measure the rate of progress over the past century or so, and the number of researchers in the field over the same period. For example, here's the transistor data:



This is the standard presentation of Moore's Law – the number of transistors you can fit on a chip doubles about every two years (eg grows by 35% per year). This is usually presented as an amazing example of modern science getting things right, and no wonder – it means you can go from a few thousand transistors per chip in 1971 to many million today, with the corresponding increase in computing power.

But BJRW have a pessimistic take. There are eighteen times more people involved in transistor-related research today than in 1971. So if in 1971 it took 1000 scientists to increase transistor density 35% per year, today it takes 18,000 scientists to do the same task. So apparently the average transistor scientist is eighteen times less productive today than fifty years ago. That should be surprising and scary.

But isn't it unfair to compare percent increase in transistors with absolute increase in transistor scientists? That is, a graph comparing absolute number of transistors per chip vs. absolute number of transistor scientists would show two similar exponential trends. Or a graph comparing percent change in transistors per year vs. percent change in number of transistor scientists per year would show two similar linear trends. Either way, there would be no problem and productivity would appear constant since 1971. Isn't that a better way to do things?

A lot of people asked paper author Michael Webb this at the conference, and his answer was no. He thinks that intuitively, each “discovery” should decrease transistor size by a certain amount. For example, if you discover a new material that allows transistors to be 5% smaller along one dimension, then you can fit 5% more transistors on your chip whether there were a hundred there before or a million. Since the relevant factor is discoveries per researcher, and each discovery is represented as a percent change in transistor size, it makes sense to compare percent change in transistor size with absolute number of researchers.

Anyway, most other measurable fields show the same pattern of constant progress in the face of exponentially increasing number of researchers. Here's BJRW's data on crop yield:



The solid and dashed lines are two different measures of crop-related research. Even though the crop-related research increases by a factor of 6-24x (depending on how it's

measured), crop yields grow at a relatively constant 1% rate for soybeans, and apparently declining 3%ish percent rate for corn.

BJRW go on to prove the same is true for whatever other scientific fields they care to measure. Measuring scientific progress is inherently difficult, but their finding of constant or log-constant progress in most areas accords with [Nintil's overview of the same topic](#), which gives us graphs like



...and dozens more like it. And even when we use data that are easy to measure and hard to fake, like number of chemical elements discovered, we get the same linearity:



Meanwhile, the increase in researchers is obvious. Not only is the population increasing (by a factor of about 2.5x in the US since 1930), but the percent of people with college degrees has quintupled over the same period. The exact numbers differ from field to field, but orders of magnitude increases are the norm. For example, the number of people publishing astronomy papers [seems to have decupled](#) over the past fifty years or so.

BJRW put all of this together into total number of researchers vs. total factor productivity of the economy, and find...



...about the same as with transistors, soybeans, and everything else. So if you take their methodology seriously, over the past ninety years, each researcher has become about 25x less productive in making discoveries that translate into economic growth.

Participants at the conference had some explanations for this, of which the ones I remember best are:

1. Only the best researchers in a field actually make progress, and the best researchers are already in a field, and probably couldn't be kept *out of* the field with barbed wire and attack dogs. If you expand a field, you will get a bunch of merely competent careerists who treat it as a 9-to-5 job. A field of 5 truly inspired geniuses and 5 competent careerists will make X progress. A field of 5 truly inspired geniuses and 500,000 competent careerists will make the same X progress. Adding further competent careerists is useless for doing anything except making graphs look more exponential, and we should stop doing it. See also [Price's Law Of Scientific Contributions](#).
2. Certain features of the modern academic system, like underpaid PhDs, interminably long postdocs, endless grant-writing drudgery, and clueless funders have lowered productivity. The 1930s academic system was indeed 25x more effective at getting researchers to actually do good research.
3. All the low-hanging fruit has already been picked. For example, element 117 was discovered by an international collaboration who got an unstable isotope of berkelium from the single accelerator in Tennessee capable of synthesizing it, shipped it to a nuclear reactor in Russia where it was attached to a titanium film, brought it to a particle accelerator in a different Russian city where it was bombarded with a custom-made exotic isotope of calcium, sent the resulting data to a global team of theorists,

and eventually found a signature indicating that element 117 had existed for a few milliseconds. Meanwhile, the first modern element discovery, that of phosphorous in the 1670s, came from [a guy looking at his own piss](#). We should not be surprised that discovering element 117 needed more people than discovering phosphorous.

Needless to say, my sympathies lean towards explanation number 3. But I worry even this isn't dismissive enough. My real objection is that constant progress in science in response to exponential increases in inputs ought to be our null hypothesis, and that it's almost inconceivable that it could ever be otherwise.

Consider a case in which we extend these graphs back to the beginning of a field. For example, psychology started with Wilhelm Wundt and a few of his friends playing around with stimulus perception. Let's say there were ten of them working for one generation, and they discovered ten revolutionary insights worthy of their own page in Intro Psychology textbooks. Okay. But now there are about a hundred thousand experimental psychologists. Should we expect them to discover a hundred thousand revolutionary insights per generation?

Or: the economic growth rate in 1930 was 2% or so. If it scaled with number of researchers, it ought to be about 50% per year today with our 25x increase in researcher number. That kind of growth would mean that the average person who made \$30,000 a year in 2000 should make \$50 million a year in 2018.

Or: in 1930, life expectancy at 65 was increasing by about two years per decade. But if that scaled with number of biomedicine researchers, that should have increased to ten years per decade by about 1955, which would mean everyone would have become immortal starting sometime during the Baby Boom, and we would currently be ruled by a deathless God-Emperor Eisenhower.

Or: the ancient Greek world had about 1% the population of the current Western world, so if the average Greek was only 10% as likely to be a scientist as the average modern, there were only 1/1000th as many Greek scientists as modern ones. But the Greeks made such great discoveries as the size of the Earth, the distance of the Earth to the sun, the prediction of eclipses, the heliocentric theory, Euclid's geometry, the nervous system, the cardiovascular system, etc, and brought technology up from the Bronze Age to the Antikythera mechanism. Even adjusting for the long time scale to which "ancient Greece" refers, are we sure that we're producing 1000x as many great discoveries as they are? If we extended BJRW's graph all the way back to Ancient Greece, adjusting for the change in researchers as civilizations rise and fall, wouldn't it keep the same shape as does for this century? Isn't the real question not "Why isn't Dwight Eisenhower immortal god-emperor of Earth?" but "Why isn't Marcus Aurelius immortal god-emperor of Earth?"

Or: what about human excellence in other fields? Shakespearean England had 1% of the population of the modern Anglosphere, and presumably even fewer than 1% of the artists. Yet it gave us Shakespeare. Are there a hundred Shakespeare-equivalents around today? This is a harder problem than it seems – Shakespeare has become so venerable with historical hindsight that maybe nobody would acknowledge a Shakespeare-level master today even if they existed – but still, a hundred Shakespeares? If we look at some measure of great works of art per era, we find past eras giving us far more than we would predict from their population relative to our own. This is very hard to judge, and I would hate to be the guy who has to decide whether Harry Potter is better or worse than the *Aeneid*. But still? A hundred Shakespeares?

Or: what about sports? Here's marathon records for the past hundred years or so:



In 1900, there were only two local marathons (eg the Boston Marathon) in the world. Today there are over 800. Also, the world population has increased by a factor of five (more than that in the East African countries that give us literally 100% of top male marathoners). Despite that, progress in marathon records has been steady or declining. Most other Olympics sports show the same pattern.

All of these lines of evidence lead me to the same conclusion: constant growth rates in response to exponentially increasing inputs is the null hypothesis. If it wasn't, we should be expecting 50% year-on-year GDP growth, easily-discovered-immortality, and the like. Nobody expected that before reading BJRW, so we shouldn't be surprised when BJRW provide a data-driven model showing it isn't happening. I realize this in itself isn't an explanation; it doesn't tell us why researchers can't maintain a constant level of output as measured in discoveries. It sounds a little like "God wouldn't design the universe that way", which is a kind of suspicious line of argument, especially for atheists. But it at least shifts us from a lens where we view the problem as "What three tweaks should we make to the graduate education system to fix this problem right now?" to one where we view it as "Why isn't Marcus Aurelius immortal?"

And through such a lens, only the "low-hanging fruits" explanation makes sense. Explanation 1 – that progress depends only on a few geniuses – isn't enough. After all, the Greece-today difference is partly based on population growth, and population growth should have produced proportionately more geniuses. Explanation 2 – that PhD programs have gotten worse – isn't enough. There would have to be a worldwide monotonic decline in every field (including sports and art) from Athens to the present day. Only Explanation 3 holds water.

I brought this up at the conference, and somebody reasonably objected – doesn't that mean science will stagnate soon? After all, we can't keep feeding it an exponentially increasing number of researchers forever. If nothing else stops us, then at some point, 100% (or the highest plausible amount) of the human population will be researchers, we can only increase as fast as population growth, and then the scientific enterprise collapses.

I answered that the Gods Of Straight Lines are more powerful than the Gods Of The Copybook Headings, so if you try to use common sense on this problem you will fail.

Imagine being a futurist in 1970 presented with Moore's Law. You scoff: "If this were to continue only 20 more years, it would mean a million transistors on a single chip! You would be able to fit an entire supercomputer in a shoebox!" But common sense was wrong and the trendline was right.

"If this were to continue only 40 more years, it would mean ten *billion* transistors per chip! You would need more transistors on a single chip than there are humans in the world! You could have computers more powerful than any today, that are too small to even see with the naked eye! You would have transistors with like a double-digit number of atoms!" But common sense was wrong and the trendline was right.

Or imagine being a futurist in ancient Greece presented with world GDP doubling time. Take the trend seriously, and in two thousand years, the future would be fifty thousand times richer. Every man would live better than the Shah of Persia! There would have to be so many people in the world you would need to tile entire countries with

cityscape, or build structures higher than the hills just to house all of them. Just to sustain itself, the world would need transportation networks orders of magnitude faster than the fastest horse. But common sense was wrong and the trendline was right.



I'm not saying that no trendline has ever changed. Moore's Law seems to be legitimately slowing down these days. [The Dark Ages shifted](#) every macrohistorical indicator for the worse, and [the Industrial Revolution shifted](#) every macrohistorical indicator for the better. Any of these sorts of things could happen again, easily. I'm just saying that "Oh, that exponential trend can't possibly continue" has a really bad track record. I do not understand the Gods Of Straight Lines, and honestly they creep me out. But I would not want to bet against them.

[Grace et al's](#) survey of AI researchers show they predict that AIs will start being able to do science in about thirty years, and will exceed the productivity of human researchers in every field shortly afterwards. Suddenly "there aren't enough humans in the entire world to do the amount of research necessary to continue this trend line" stops sounding so compelling.

At the end of the conference, the moderator asked how many people thought that it was possible for a concerted effort by ourselves and our institutions to "fix" the "problem" indicated by BJRW's trends. Almost the entire room raised their hands. Everyone there was smarter and more prestigious than I was (also richer, and in many cases way more attractive), but with all due respect I worry they are insane. This is kind of how I imagine their worldview looking:



I realize I'm being fatalistic here. Doesn't my position imply that the scientists at Intel should give up and let the Gods Of Straight Lines do the work? Or at least that the head of the National Academy of Sciences should do something like that? That Francis Bacon was wasting his time by inventing the scientific method, and Fred Terman was wasting his time by organizing Silicon Valley? Or perhaps that the Gods Of Straight Lines were acting *through* Bacon and Terman, and they had no choice in their actions? How do we know that the Gods aren't acting through our conference? Or that our studying these things isn't the only thing that keeps the straight lines going?

I don't know. I can think of some interesting models – one made up of a thousand random coin flips a year has some nice qualities – but I don't know.

I do know you should be careful what you wish for. If you "solved" this "problem" in classical Athens, Attila the Hun would have had nukes. Remember Yudkowsky's Law of Mad Science: "Every eighteen months, the minimum IQ necessary to destroy the world drops by one point." Do you really want to make that number ten points? A hundred? I am kind of okay with the function mapping number of researchers to output that we have right now, thank you very much.



The conference was organized by Patrick Collison and Michael Nielsen; they have written up some of their thoughts [here](#).

Incorrect hypotheses point to correct observations

1. The Consciousness Researcher and Out-Of-Body Experiences

In his book [*Consciousness and the Brain*](#), cognitive neuroscientist Stanislas Dehaene writes about scientifically investigating people's reports of their out-of-body experiences:

... the Swiss neurologist Olaf Blanke[did a] beautiful series of experiments on out-of-body experiences. Surgery patients occasionally report leaving their bodies during anesthesia. They describe an irrepressible feeling of hovering at the ceiling and even looking down at their inert body from up there. [...]

What kind of brain representation, Blanke asked, underlies our adoption of a specific point of view on the external world? How does the brain assess the body's location? After investigating many neurological and surgery patients, Blanke discovered that a cortical region in the right temporoparietal junction, when impaired or electrically perturbed, repeatedly caused a sensation of out-of-body transportation. This region is situated in a high-level zone where multiple signals converge: those arising from vision; from the somatosensory and kinesthetic systems (our brain's map of bodily touch, muscular, and action signals); and from the vestibular system (the biological inertial platform, located in our inner ear, which monitors our head movements). By piecing together these various clues, the brain generates an integrated representation of the body's location relative to its environment. However, this process can go awry if the signals disagree or become ambiguous as a result of brain damage. Out-of-body flight "really" happens, then—it is a real physical event, but only in the patient's brain and, as a result, in his subjective experience. The out-of-body state is, by and large, an exacerbated form of the dizziness that we all experience when our vision disagrees with our vestibular system, as on a rocking boat.

Blanke went on to show that any human can leave her body: he created just the right amount of stimulation, via synchronized but delocalized visual and touch signals, to elicit an out-of-body experience in the normal brain. Using a clever robot, he even managed to re-create the illusion in a magnetic resonance imager. And while the scanned person experienced the illusion, her brain lit up in the temporoparietal junction—very close to where the patient's lesions were located.

We still do not know exactly how this region works to generate a feeling of self-location. Still, the amazing story of how the out-of-body state moved from parapsychological curiosity to mainstream neuroscience gives a message of hope. Even outlandish subjective phenomena can be traced back to their neural origins. The key is to treat such introspections with just the right amount of seriousness. They do not give direct insights into our brain's inner mechanisms; rather, they constitute the raw material on which a solid science of consciousness can be properly founded.

The naive hypotheses that out-of-body experiences represented the spirit genuinely leaving the body, were incorrect. But they were still pointing to a real observation,

namely that there are conditions which create a subjective experience of leaving the body. That observation could then be investigated through scientific means.

2. The Artist and the Criticism

In art circles, there's a common piece of advice that goes along the lines of:

When people say that they don't like something about your work, you should treat that as valid information.

When people say *why* they don't like it or what you could do to fix it, you should treat that with some skepticism.

Outside the art context, if someone tells you that they're pissed off with you as a person (or that you make them feel good), then that's likely to be true; but the reason that they give you may not be the true reason.

People [have poor introspective access](#) to the reasons why they like or dislike something; when they are asked for an explanation, they often literally fabricate their reasons. Their explanation is likely false, even though it's still pointing to *something* in the work having made them dislike it.

3. The Traditionalist and the Anthropologist

The Scholar's Stage blog post "[Tradition is Smarter Than You Are](#)", quotes Joseph Henrich's [The Secret of Our Success](#) which reports that many folk traditions, such as not eating particular fish during pregnancy, are adaptive: not eating that fish during pregnancy is good for the child, mother, or both. But the people in question often do not know *why* they follow that tradition:

We looked for a shared underlying mental model of why one would not eat these marine species during pregnancy or breastfeeding—a causal model or set of reasoned principles. Unlike the highly consistent answers on what not to eat and when, women's responses to our why questions were all over the map. Many women simply said they did not know and clearly thought it was an odd question. Others said it was "custom." Some did suggest that the consumption of at least some of the species might result in harmful effects to the fetus, but what precisely would happen to the fetus varied greatly, though a nontrivial segment of the women explained that babies would be born with rough skin if sharks were eaten and smelly joints if morays were eaten. Unlike most of our interview questions on this topic, the answers here had the flavor of post-hoc rationalization: "Since I'm being asked for a reason, there must be a reason, so I'll think one up now." This is extremely common in ethnographic fieldwork, and I've personally experienced it in the Peruvian Amazon with the Matsigenka and with the Mapuche in southern Chile.

The people's hypotheses for why they do something is wrong. But their behavior is still pointing to the fish in question being bad to eat during pregnancy.

4. The Martial Artist and the Ki

In [Types of Knowing](#), Valentine writes:

Another example is the "[unbendable arm](#)" in martial arts. I learned this as a matter of "[extending ki](#)": if you let magical life-energy blast out your fingertips,

then your arm becomes hard to bend much like it's hard to bend a hose with water blasting out of it. This is obviously not what's really happening, but thinking this way often gets people to be able to do it after a few cumulative hours of practice.

But you know what helps better?

Knowing the physics.

Turns out that the unbendable arm is a leverage trick: if you treat the upward pressure on the wrist as a fulcrum and you push your hand down (or rather, raise your elbow a bit), you can redirect that force and the force that's downward on your elbow into each other. Then you don't need to be strong relative to how hard your partner is pushing on your elbow; you just need to be strong enough to redirect the forces into each other.

Knowing this, I can teach someone to pretty reliably do the unbendable arm in under ten minutes. No mystical philosophy needed.

The explanation about magical life energy was false, but it was still pointing to a useful trick that could be learned and put to good use.

Observations and the hypotheses developed to explain them often get wrapped up, causing us to evaluate both as a whole. In some cases, we only hear the hypothesis rather than the observation which prompted it. But people usually don't pull their hypotheses out of entirely thin air; even an incorrect hypothesis is usually [entangled with](#) some correct observations. If we can isolate the observation that prompted the hypothesis, then we can treat the hypothesis as a [burdensome detail](#) to be evaluated on its own merits, separate from the original observation. At the very least, the existence of an incorrect but common hypothesis suggests to us that there's *something* going on that needs to be explained.

Cross-posted
[Bibliography](#)

[Learning](#)

[Environment](#)

[Agent](#)

[Successor](#)

[Intervention](#)

[Halting](#)

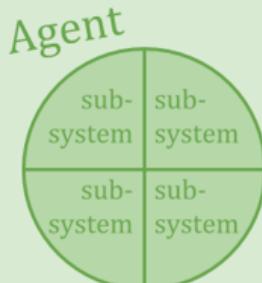
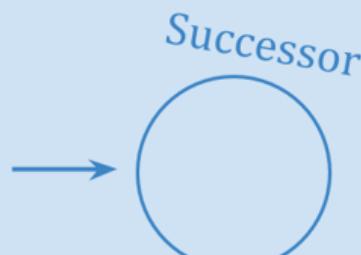
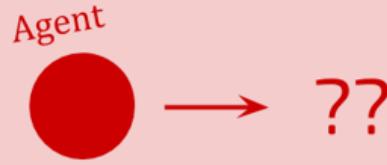
[Goodhart](#)

[B-halting?](#)

[Halting??](#)

[A](#)

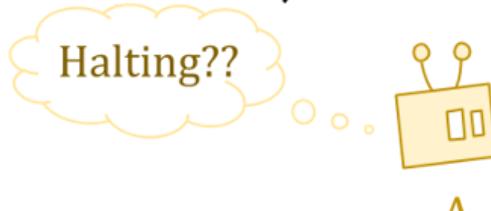
[Halting](#)



[Halting??](#)

[Causal Goodhart](#)

[T](#) [J](#) [BEST](#)



[A](#)

[Halting??](#)



even
more
money.

Counterintuitive Comparative Advantage

This has been sitting in my drafts folder since 2011. Decided to post it today given the [recent post about Dunning–Kruger](#) and related discussions.

The standard rationalist answer when someone asks for career advice is "find your comparative advantage." I don't have any really good suggestions about how to make this easier, but it seems like a good topic to bring up for discussion.

If 15 years ago (when I was still in college and my initial career choice hadn't been finalized yet), someone told me that perhaps I ought to consider a career in philosophy, I would have laughed. "You must be joking. *Obviously*, I'll be really bad at doing philosophy," I would have answered. I thought of myself as a natural born programmer, and that's the career direction I ended up choosing.

As it turns out, I am a pretty good programmer, and a terrible philosopher, but it also happens to be the case that just about everyone else is even *worse* at doing philosophy, and getting some philosophical questions right might be *really* important.

The usual (instinctive) way for someone to choose a career is probably to pick a field that they think they will be particularly good at, using a single standard of goodness across all of the candidate fields. For example, the implicit reasoning behind my own career choice could be something like "Given a typical programming problem, I can solve it in a few hours with high probability. Whereas, given a typical philosophical problem, I can at best solve it after many years with low probability."

On the other hand, comparative advantage says that in addition to your own abilities, you should also consider how good other people are (or will be) at various fields, and how valuable the outputs of those fields are (will be). Unless you're only interested in maximizing income, and the fields you're considering are likely to remain stable over your lifetime (in which case you can just compare current salaries, although apparently many people don't even do that), this can be pretty tricky.

(There doesn't appear to be any previous OB/LW posts on comparative advantage. The closest I could find is Eliezer's [Money: The Unit of Caring](#). Most discussions elsewhere seem to focus on simple static examples where finding comparative advantage is relatively trivial.)

Today (in 2018) there's an [80,000 Hour article](#) about comparative advantage but that is more about how to find one's comparative advantage in a community of people who share a cause, like in EA, rather in the wider economy.

I would also add (in 2018) that besides everyone else lacking skill or talent at something, an even bigger source of comparative advantage is being one of the first people to realize that a problem is a problem, or to realize an important new variant or subproblem of an existing problem. In that case, everyone else is really bad at solving that problem just because they have no idea the problem even exists.

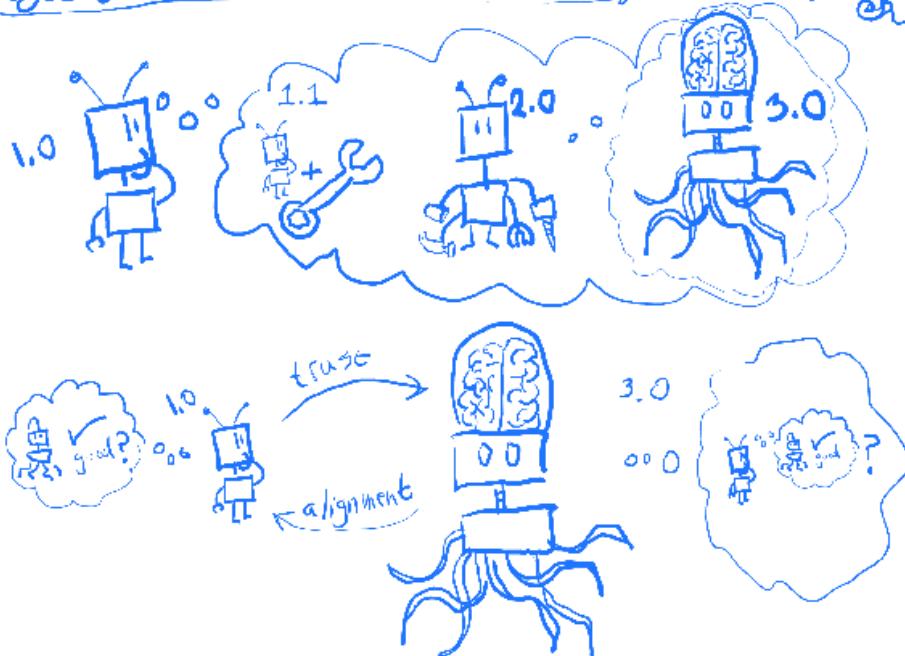
Robust Delegation

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

(A longer text-based version of this post is also available on MIRI's blog [here](#), and the bibliography for the whole sequence can be found [here](#))

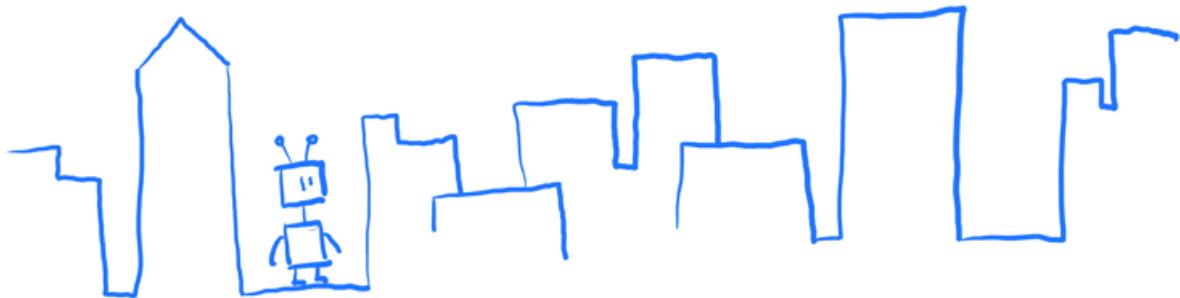
[Robust Delegation]

agent can reason about itself and improve



[Abram Demski and Scott Garrabrant]

Because the world is big, the agent as it is may be inadequate to accomplish its goals, including in its ability to think.



Because the agent is made of parts, it can improve itself and become more

... improve ... - - - - -
capable.

Improvements can take many forms:



tools



successor
agents



just learning
and growing
over time

However, the successors/tools need to be more capable for this to be worthwhile.

This gives rise to a special type of principal/agent problem: how can you trust something as complicated as, or more complicated than, you?

principal:  human

 AI

 AI

agent:  AI

 Same
 later

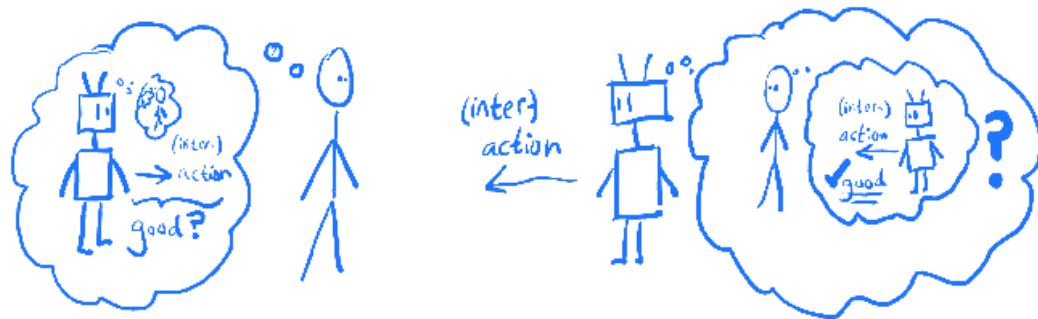
 AI
smarter AI
built by first

Alignment
Problem

Tiling
Agent
Problem

reflective stability
of goal system under
self-improvement

The problem is not (just) that the successor agent might be malicious. The problem is that we don't even know what it means not to be.



This problem seems hard from both points of view.

I want to emphasize that the view in which there are multiple forms of the problem is a dualistic view.

To an embedded agent, the future self is not privileged; it is just another part of the environment. There is no difference between making a successor and preserving your own goals.

So, when we talk about designing agents which are helpful to humans, it is not just about that. It is about the fundamental problem of being an agent that persists and learns over time.

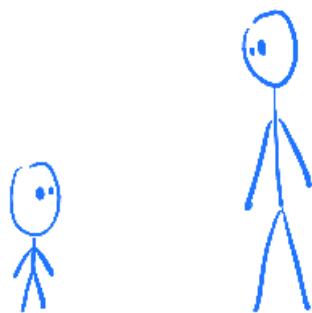
We call this cluster of problems

Robust Delegation:

- Vingeian reflection
- Tiling problem
- Averting Goodhart's Law
- Value Loading
- Corrigibility
- Informed Oversight

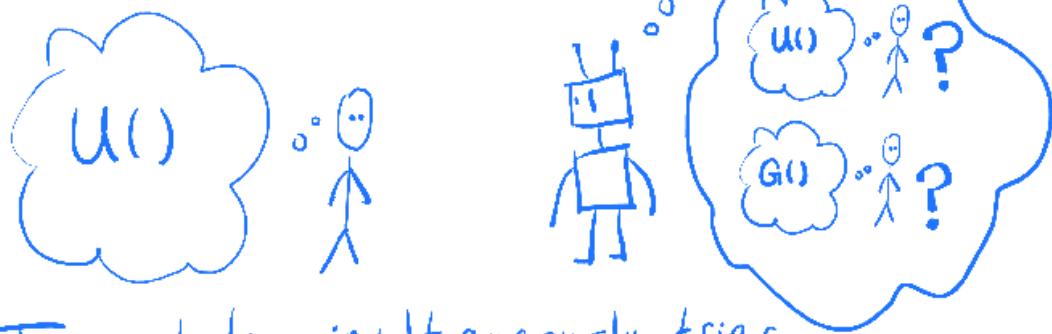
Vingeian Reflection

Imagine you are playing the CIRL game
with a toddler:



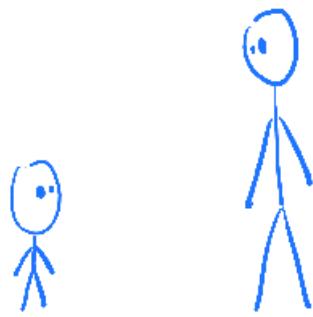
CIRL means Cooperative Inverse Reinforcement learning.

The idea behind CIRL is to define what it means for a robot to collaborate with a human.



The robot simultaneously tries to infer what the human wants while helping.

So, what if you're trying to help someone who is very confused about the universe?

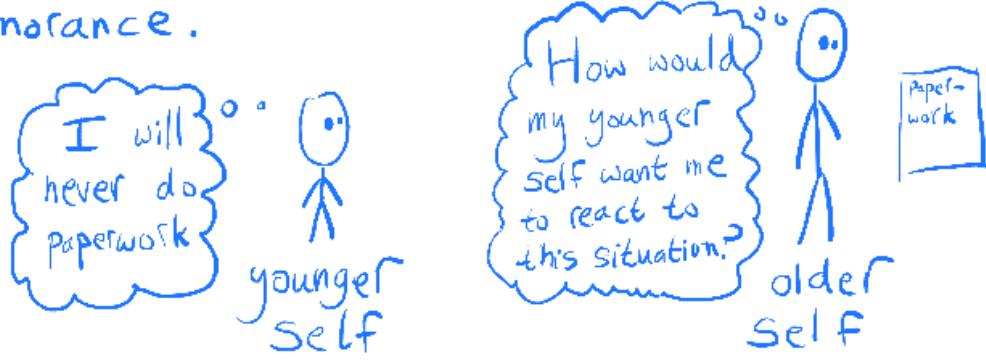


- From your standpoint, the toddler may be too irrational to be seen as optimizing anything.
- The toddler may have an ontology in which it is optimizing something, but you can see that ontology doesn't make sense.
- Maybe you notice that if you set up questions in the right way, you can make the toddler seem to want almost anything.

Part of the problem is that the "helping" agent has to be bigger in some sense in order to be more capable; but, this seems to imply that the "helped" agent can't be a very good supervisor for the "helper".



For example, updateless decision theory eliminates dynamic inconsistencies in decision theory by, rather than maximizing expected utility of your action given what you know, maximizing expected utility of reactions to observations, from a state of ignorance.



Appealing as this may be as a way to achieve reflective consistency, it creates a strange situation in terms of computational complexity:

If actions are type A, and observations are type O, reactions to observations are type $O \rightarrow A$ -- a much larger space to optimize over than A alone. And we're expecting our smaller self to be able to do that!

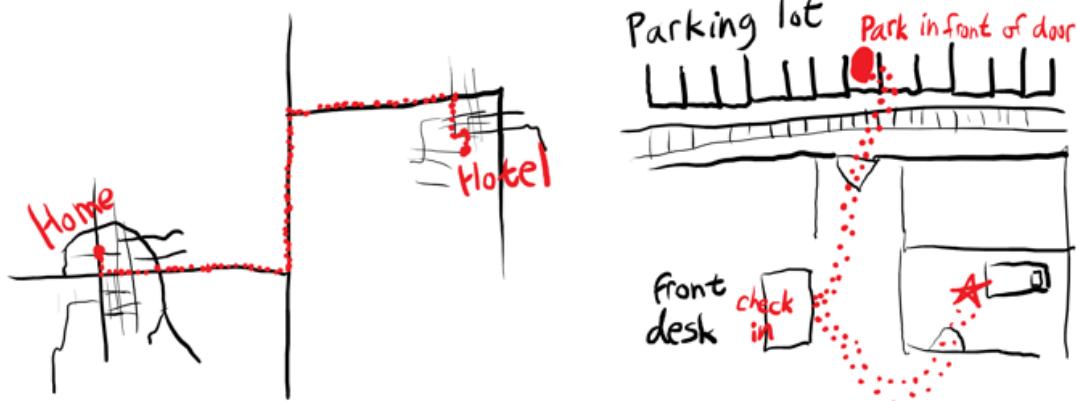
This seems bad.

One way to more crisply state the problem is:

We should be able to trust that our future self is applying its intelligence to the pursuit of our goals without being able to predict precisely what our future self will do.

This criterion is called Vingean reflection.

For example, you might plan your driving route before visiting a new city, but you do not plan your steps. You plan to some level of detail, and trust that your future self can figure out the rest.



Vingeian Reflection is difficult to examine via classical Bayesian decision theory because logical omniscience is assumed, so that the assumption that the agent knows future actions are rational is synonymous with the assumption that the agent knows its future self will act according to one particular optimal policy which the agent can predict in advance.

We have some limited models of Vingean reflection (see Tiling Agents for Self-Modifying AI, and the Löbian Obstacle by Yudkowsky & Herreshoff). A successful approach must walk the narrow line between two problems:

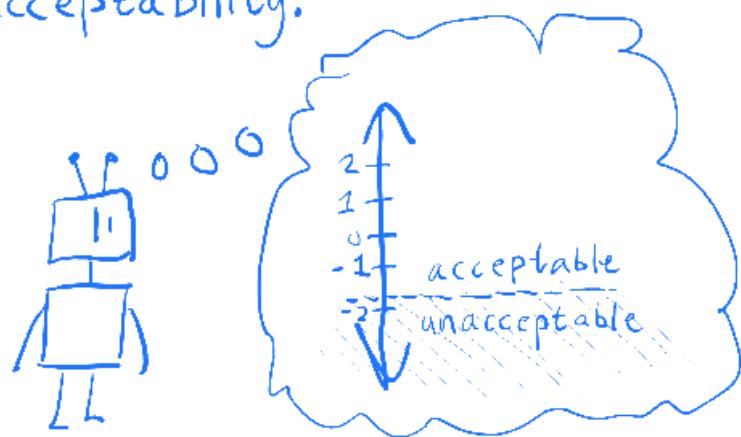
Agents who trust their future self because they trust the output of their own reasoning are inconsistent

Löbian Obstacle

Agents who trust their future selves without reason tend to be consistent but unsound & untrustworthy, and will put off tasks forever because they can do it later

Procrastination Paradox

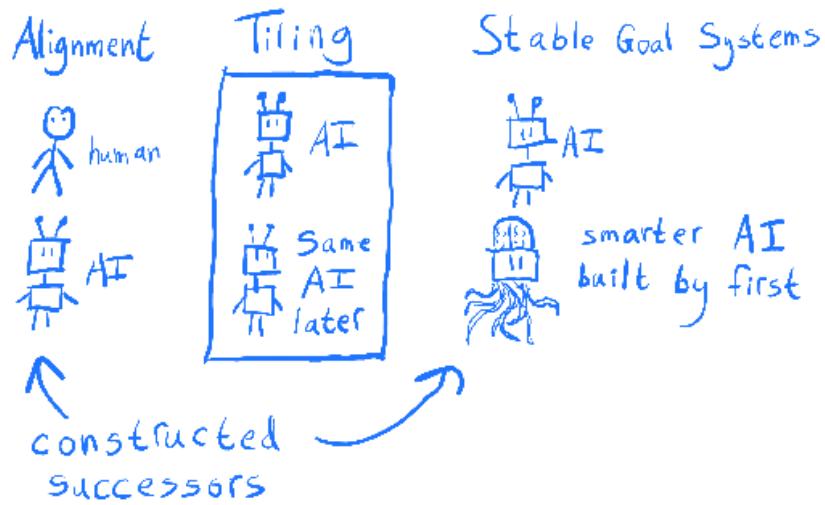
The Vingean reflection results so far apply only to limited sorts of decision procedures, such as satisficers aiming for a threshold of acceptability.



Averting Goodhart's Law

So, there is plenty of room for improvement, getting tiling results for more useful decision procedures and under weaker assumptions.

However, there is more to the robust delegation problem than just tiling and Vingean reflection.



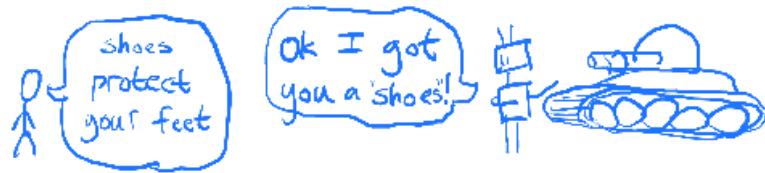
When you construct another agent, rather than delegating to your future self, you more directly face a problem of value loading.

Two facts conspire against us:

→ We don't know what we want.



→ Optimization amplifies slight differences between what we say we want and what we really want.

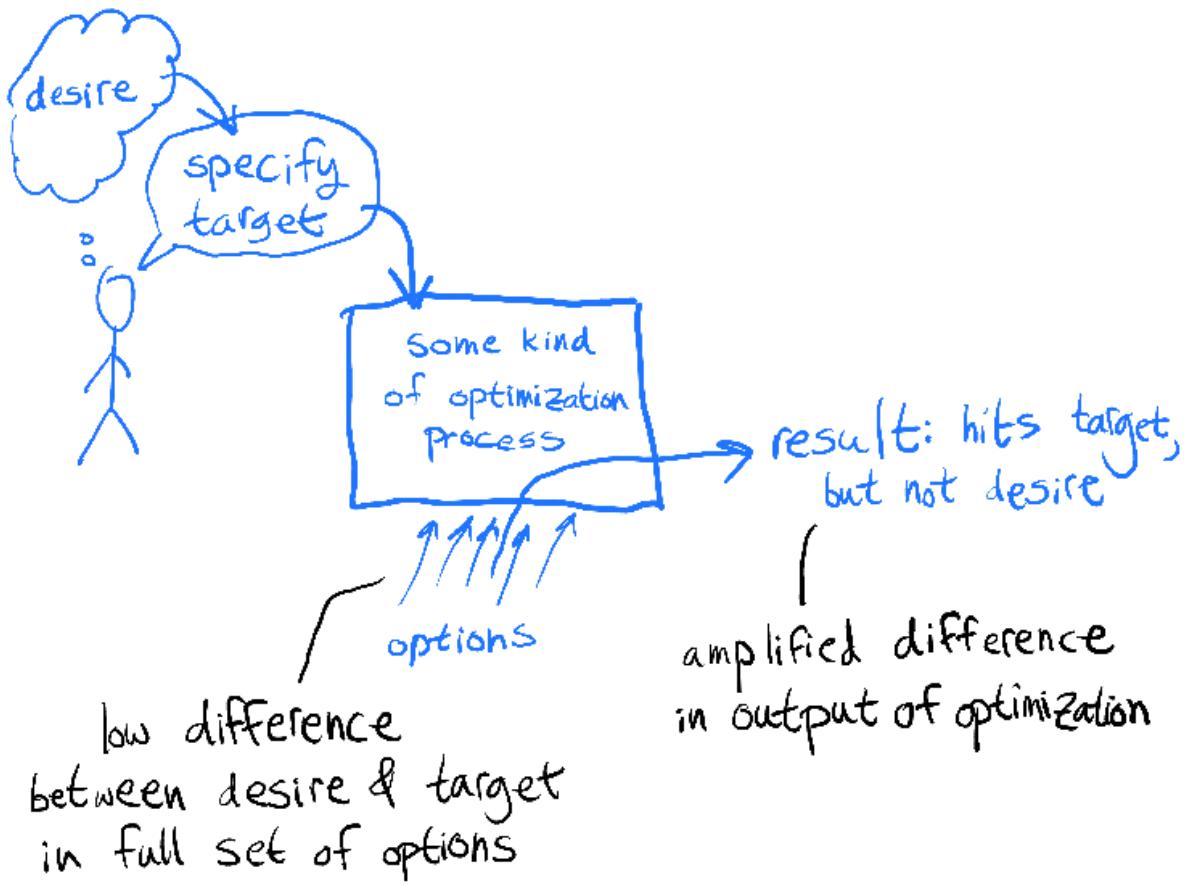


The misspecification-amplifying effect
is known as Goodhart's Law:

"Any observed statistical
regularity will tend to collapse
once pressure is placed upon
it for control purposes."

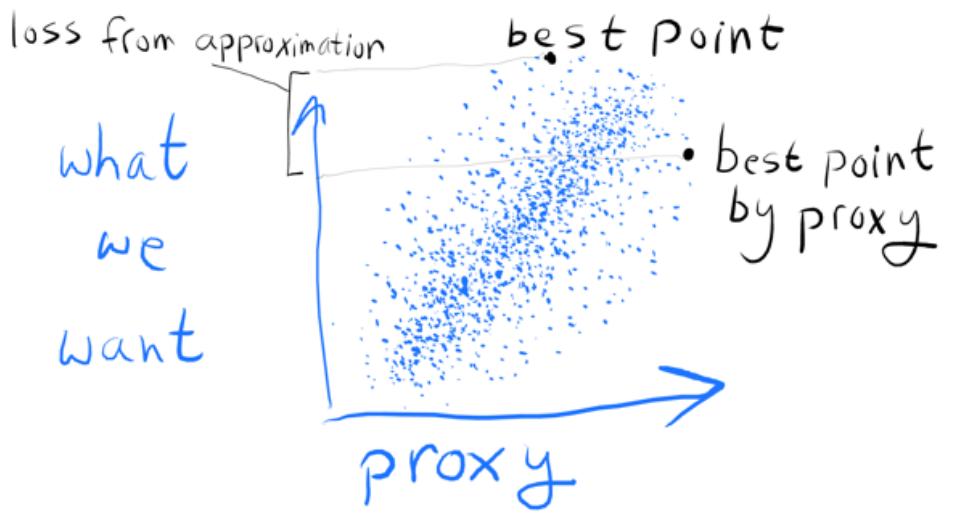
— Charles Goodhart

IE: when we specify a target for optimization, it is reasonable to expect it to be correlated with what we want (perhaps highly correlated, even); but, unfortunately, this does not mean that optimizing it will get us closer to what we want -- especially at high levels of optimization.



There ^(at least) are four types of Goodhart:

- regressional
- extremal
- causal
- adversarial



regressional

Regressional Goodhart can occur when there is a less than perfect correlation between the proxy and the goal.

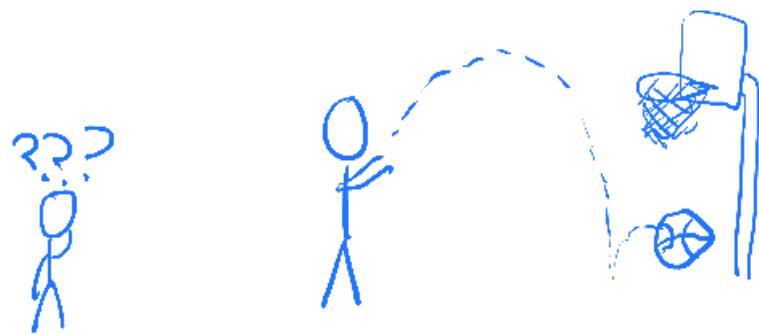
It is more commonly known as the optimizer's curse, and it is related to regression to the mean.

For example, you might draft players for a basketball team based on height alone.



This isn't a perfect heuristic, but there is a correlation between height and basketball ability, which you can utilize to make your selection.

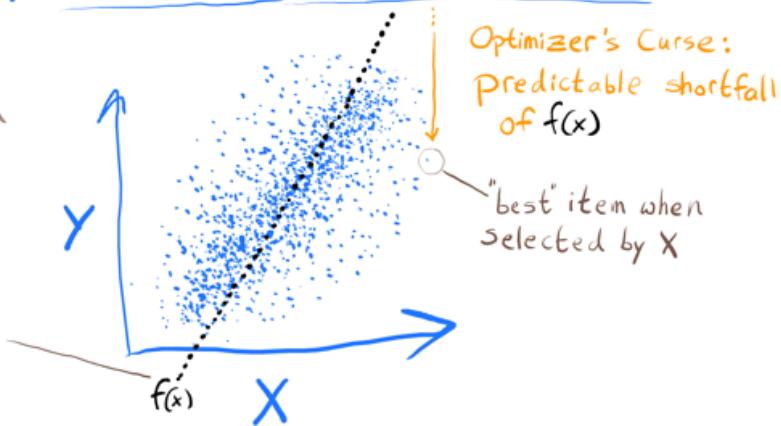
It turns out that, in a certain sense,
you will be predictably disappointed
if you expect the general trend to hold
up as strongly for your selected team.



Stated in statistical terms: an unbiased estimate of Y given X is not an unbiased estimate of Y when we select for the best X .

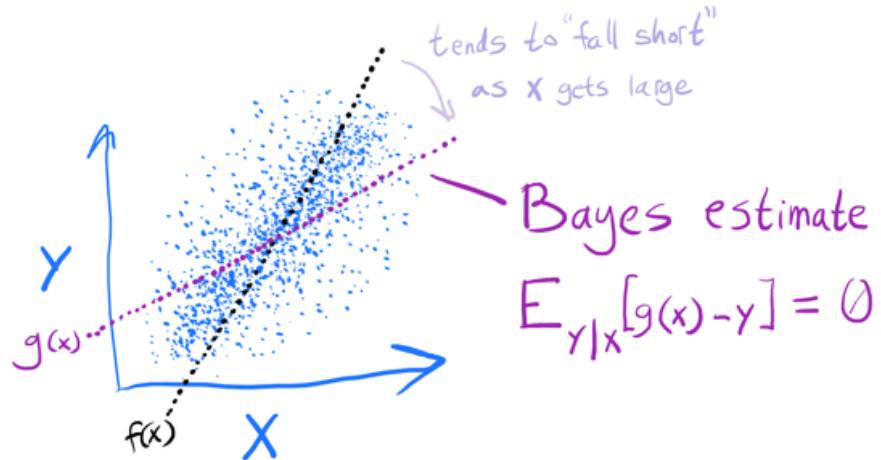
$f(x)$ is an unbiased estimate of y as a function of x , ie, $f(x)$ satisfies

$$E_{x|y} [f(x) - y] = 0$$

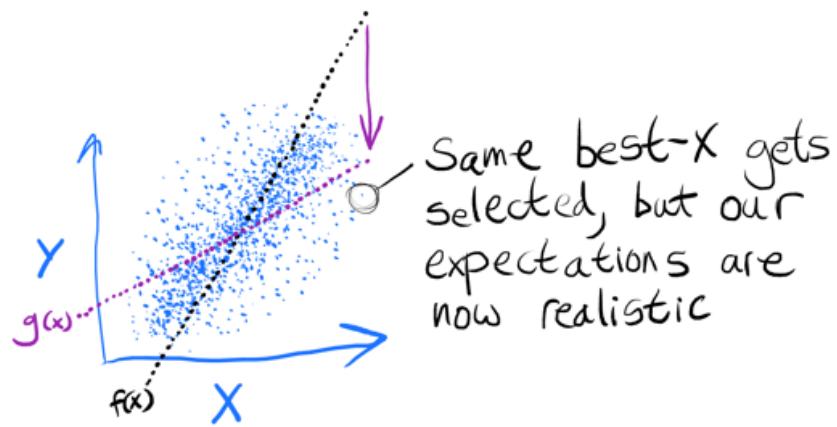


In that sense, we can expect to be disappointed when we use X as a proxy of Y for optimization purposes.

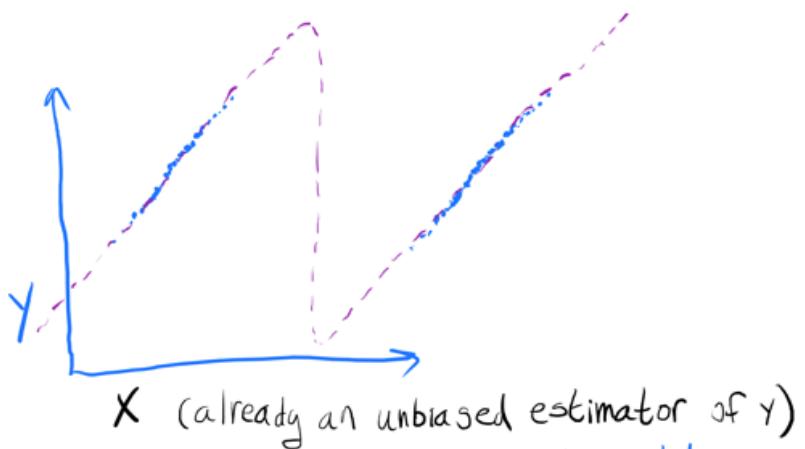
Using a Bayes estimate instead of an unbiased estimate, we can eliminate this sort of predictable disappointment.



The Bayes estimate accounts for the noise in X , bending toward typical Y -values.



This doesn't necessarily allow us to get a better Y value, since we still only have the information content of X to work with. However, it sometimes may.

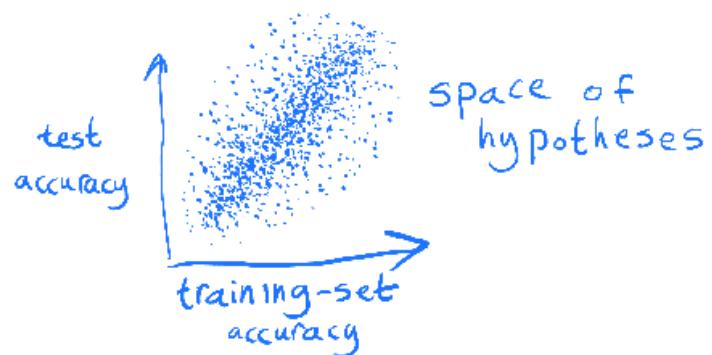


If Y is normally distributed with variance 1, and X is $Y \pm 10$ with even odds of + or -, a Bayes estimate will give better optimization results by almost entirely removing the noise.

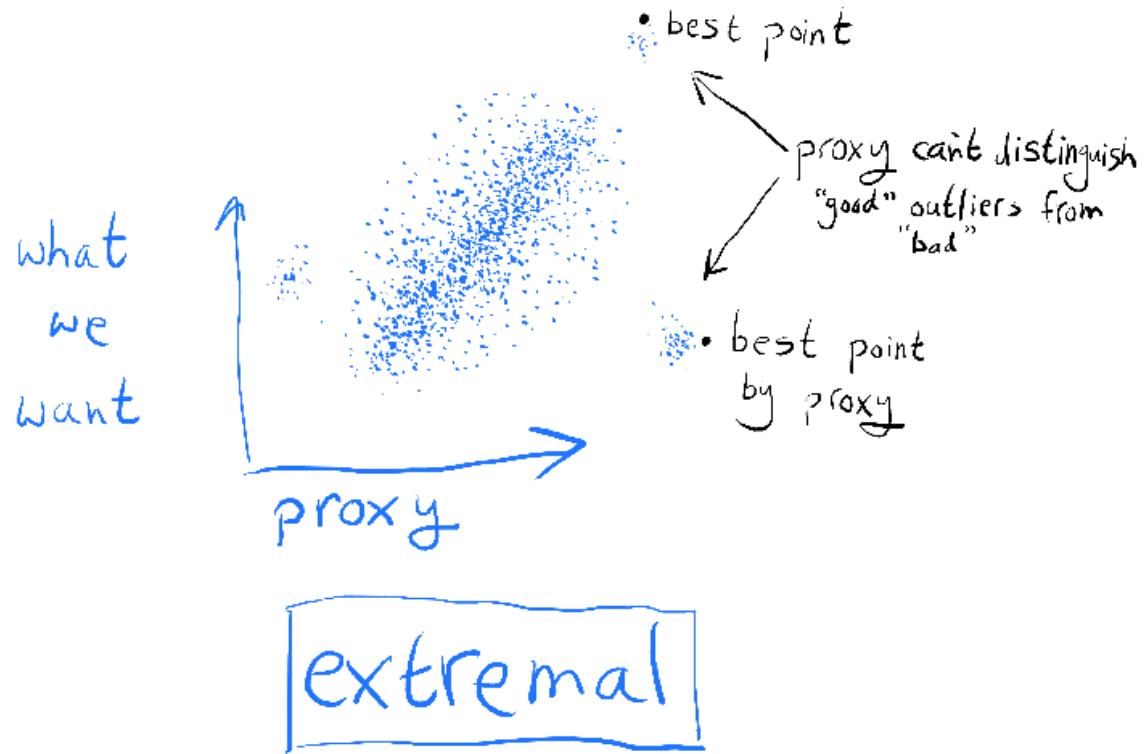
Regressional Goodhart seems like the easiest form of Goodhart to beat: just use Bayes! However, there are two big problems with this solution.

- Bayesian estimators are very often intractable in cases of interest.
- It only makes sense to trust the Bayes estimate under a **realizability** assumption.

A case where both of those problems become critical is computational learning theory.



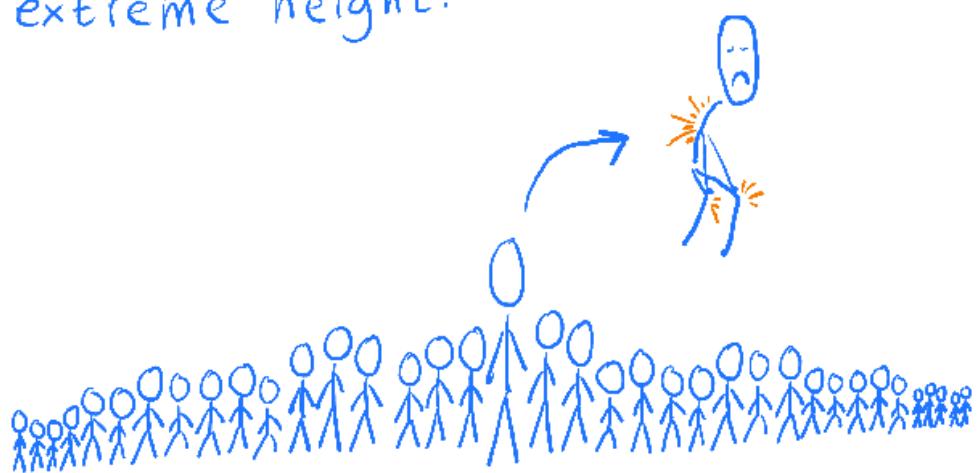
It often isn't computationally feasible to calculate the Bayesian expected generalization error of a hypothesis. But, even if you could, you still would wonder whether your chosen prior reflected the world well enough.



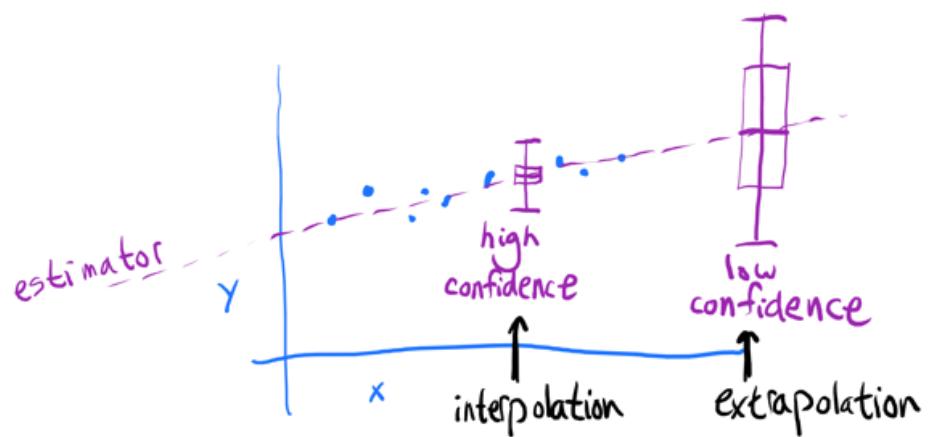
In extremal Goodhart, optimization pushes you outside the range where the correlation exists, into portions of the distribution which behave very differently.



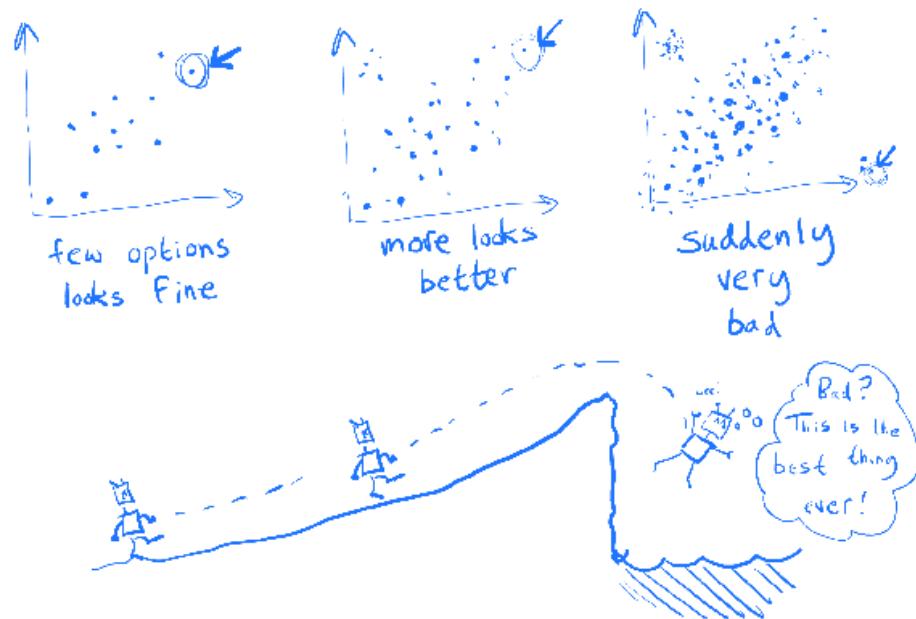
For example, if you select the tallest people in the entire world to play basketball, they may do quite poorly due to health problems which come with extreme height.



The difference between this and regressive Goodhart is related to the classical interpolation/extrapolation distinction.



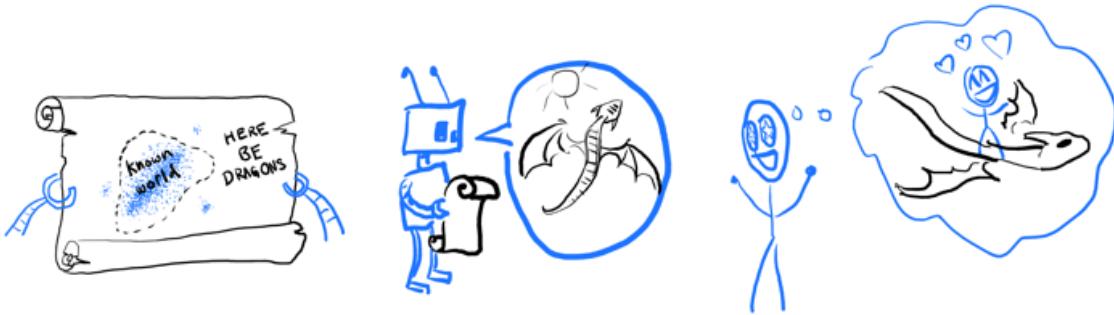
Unlike regressive Goodhart, this involves a sharp change in behavior as the system is scaled up, making it more difficult to anticipate.



As in the regressional case, a Bayesian solution addresses this concern in principle, if you trust a probability distribution to reflect the possible risks sufficiently well.

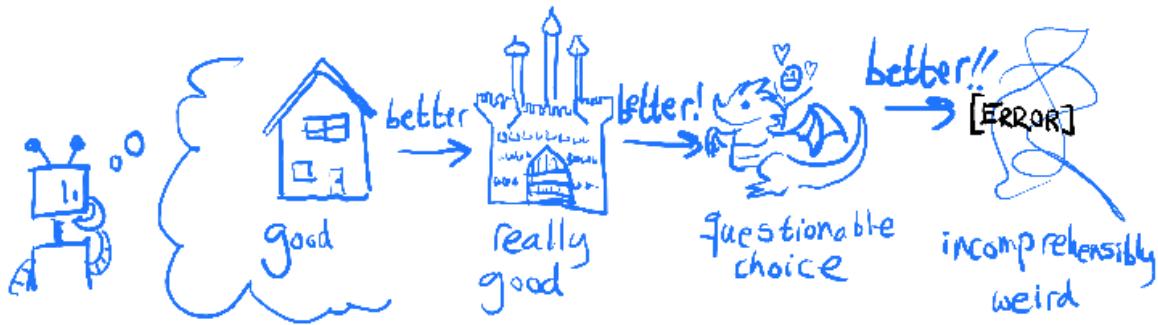
However, the realizability concern seems even more prominent here. Can a prior be trusted to anticipate problems with proposals highly optimized to look good to that specific prior?

Certainly a human's judgement couldn't be trusted under such conditions...



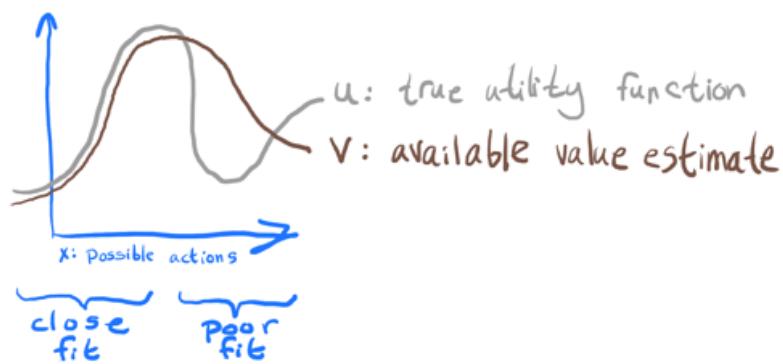
... an observation which suggests that this problem remains even if a system's judgements about value perfectly reflect a human's.

We might say the problem is that "typical" outputs avoid extremal Goodhart but "optimizing too hard" takes you out of the realm of the typical — but how can we formalize "optimizing too hard" in decision-theoretic terms?



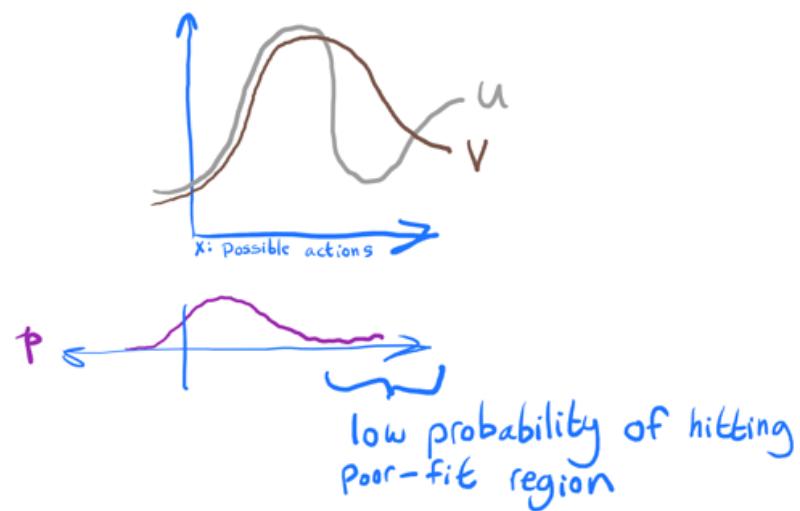
Quantilization offers a formalization.

Imagine a proxy $v(x)$ as a "corrupted" version of a true function $u(x)$.

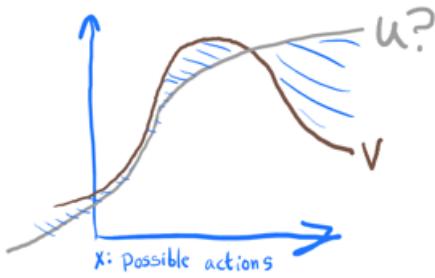


There might be different regions where the corruption is better or worse.

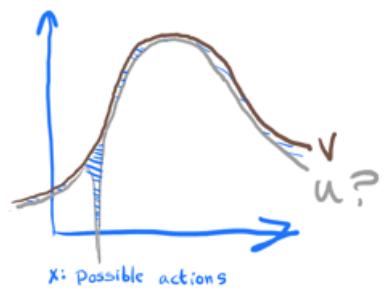
Now suppose that we can specify a "trusted" probability distribution $P(x)$, for which we are confident that the average error is below some threshold c .



By stipulating P and c , we give information about where to find low-error points without any estimates of u or of the actual error at any one point.



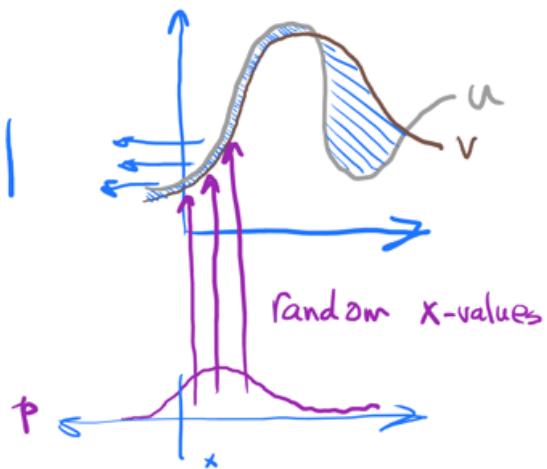
growing error concentrated toward edges of P ?



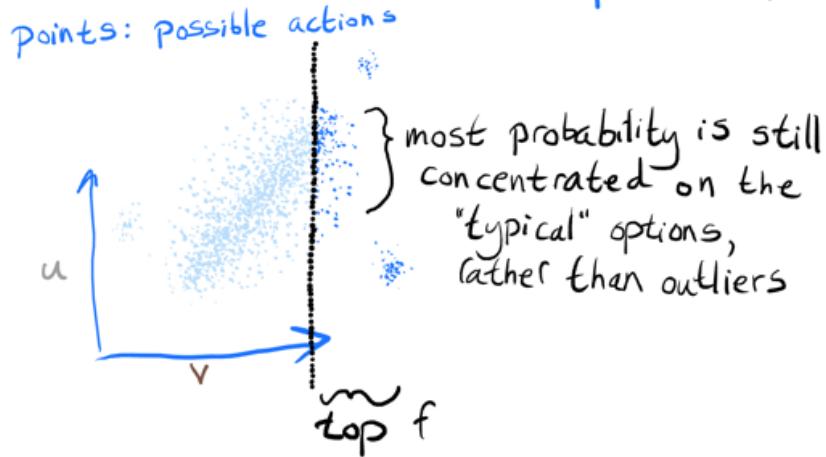
a small but severe spike of error right in the middle of P ?

When we select actions randomly from P , we can be sure that there is a low probability of high error regardless.

$$c \geq \mathbb{E}_{x \sim P} |v(x) - u(x)|$$

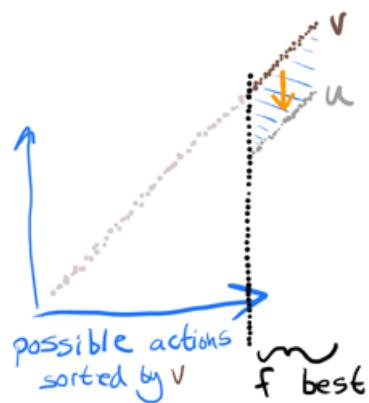


So, how do we use this to optimize?



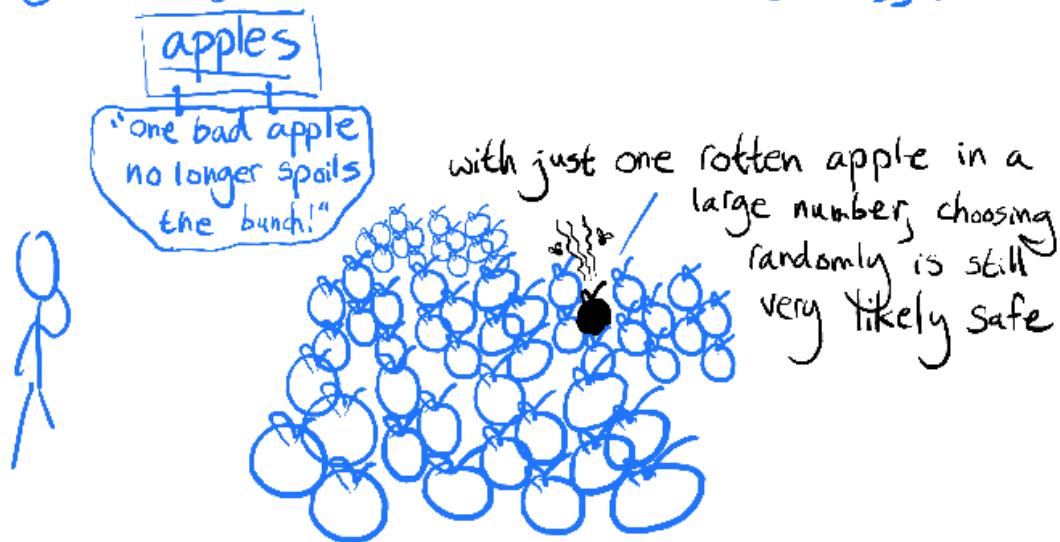
A quantilizer selects from P , but discarding all but the top fraction f ; for example, the top 1%.

This guarantees at most $\frac{c}{f}$ expected overestimation of value, since in the worst case, all of the error was over-estimation of the f_{best} .



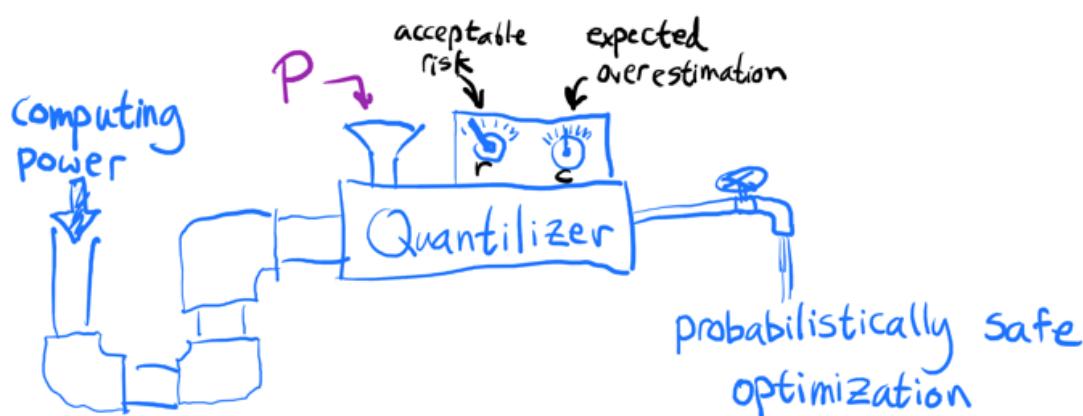
We can therefore choose an acceptable risk, $r = \frac{c}{f}$, and set the parameter f as c/r .

Quantilization is in some ways very appealing, since it allows us to specify safe classes of actions without trusting every, or any, individual action in the class.

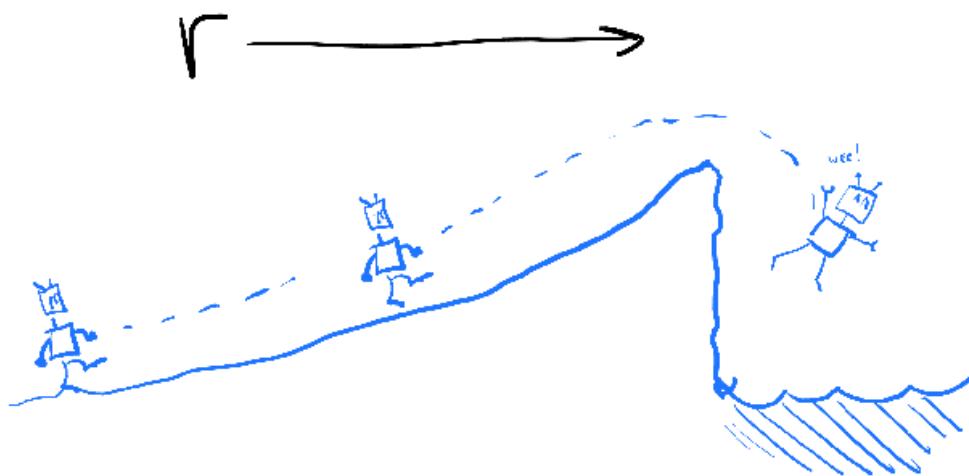


However, it also leaves much to be desired.

Where do "trusted" distributions come from?
How do you estimate expected error c ,
or select acceptable risk r ?



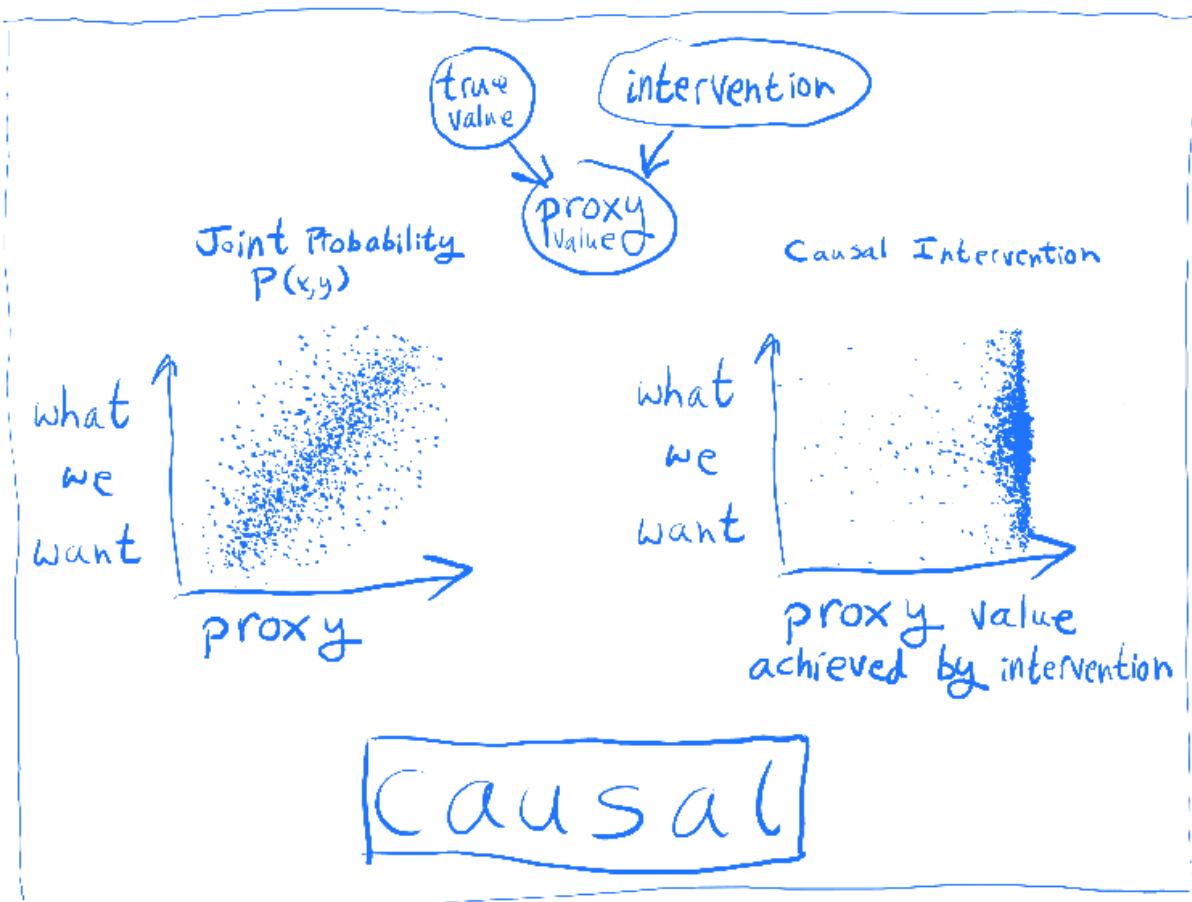
This is especially bad because it gives you a knob to turn which will apparently improve performance while increasing risk, until (possibly sudden) failure.



Also, quantilization doesn't seem likely to tile:
a quantilizing agent has no special reason
to preserve the quantilization algorithm
when making self-improvements or new agents.



So, all told, it seems like there is
room for improvement in the handling
of extremal Goodhart.



Another way optimization can go wrong is by using a proxy which more directly falls apart when we use it to optimize: the act of selecting for it breaks the connection to what we care about.

For example, you might play basketball in order to become tall.

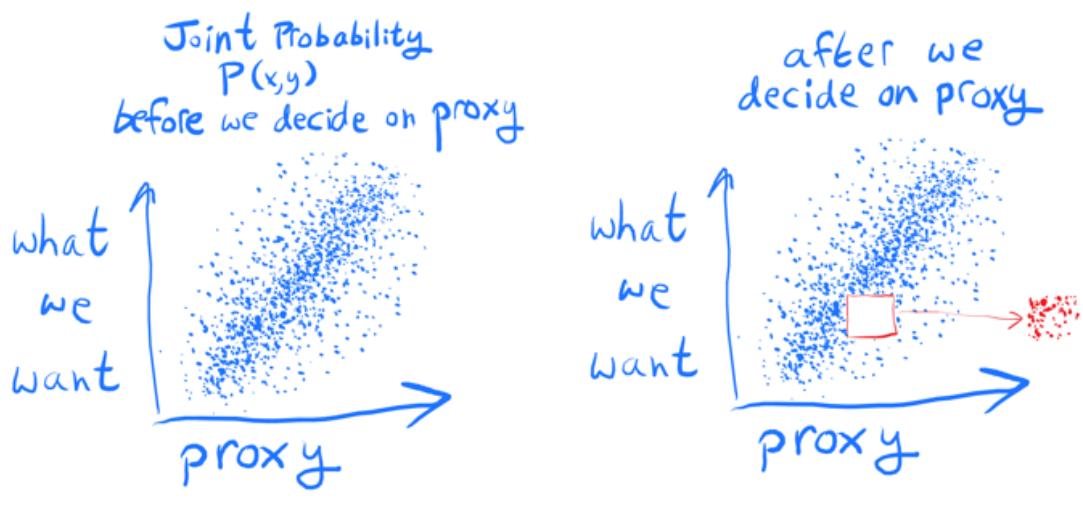


The only way to avoid this sort of mistake is to get counterfactuals right.

This might seem like punting to decision theory, but the connection here enriches both: counterfactuals have to address questions of trust due to tiling concerns, and, trust has to address counterfactual concerns due to causal Goodhart.

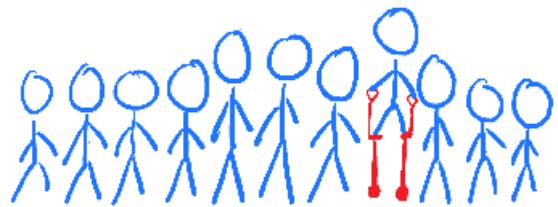
Yet again, one of the big challenges here is realizability.

As mentioned in embedded world-models, even if you have the right theory of how counterfactuals work generally, Bayesian learning doesn't provide much of a guarantee of learning to select actions well without assuming realizability.



adversarial

Finally, there is adversarial Goodhart, in which agents actively make our proxy worse by intelligently manipulating it.



For example, if applicants to your basketball team know how you are choosing players, some will specifically practice on what you check, neglecting other aspects of the game.

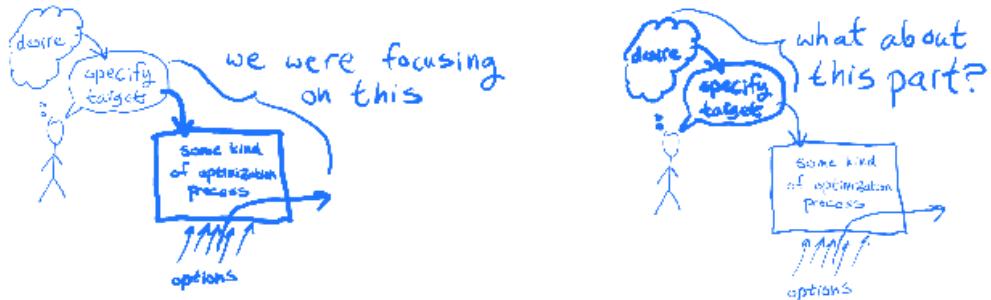
This category is the most common interpretation of Goodhart's remark. At first, it may seem less relevant to our concerns here. However, when searching in a large space which is sufficiently rich, there are bound to be some elements of that space which implement adversarial strategies.

But, that is a subject for the subsystem alignment section.

Value Loading & Value Learning

Remember none of this would happen if a system were optimizing what we wanted directly, rather than optimizing a proxy.

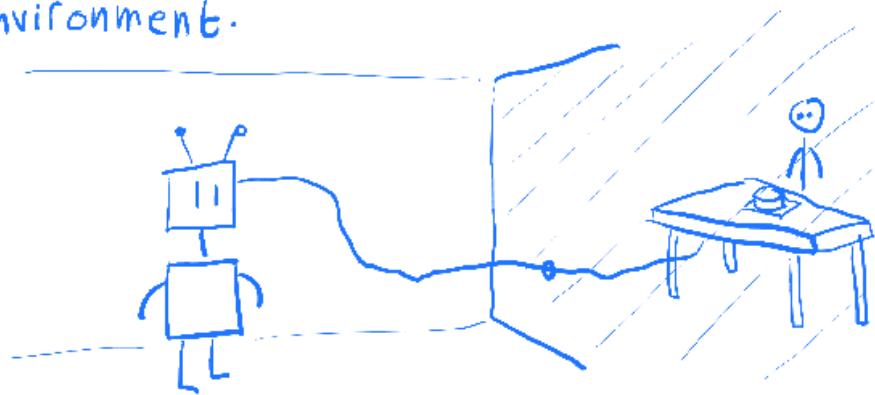
Besides anti-Goodhart measures, it would obviously help to be able to specify what we want precisely.



Unfortunately, this is hard. So, can a successor agent help us with this? If it is very intelligent, maybe it can learn what we want?



AIXI learns what to do through a reward signal which it gets from the environment.



We can imagine humans have a button which they press when AIXI does something they like.

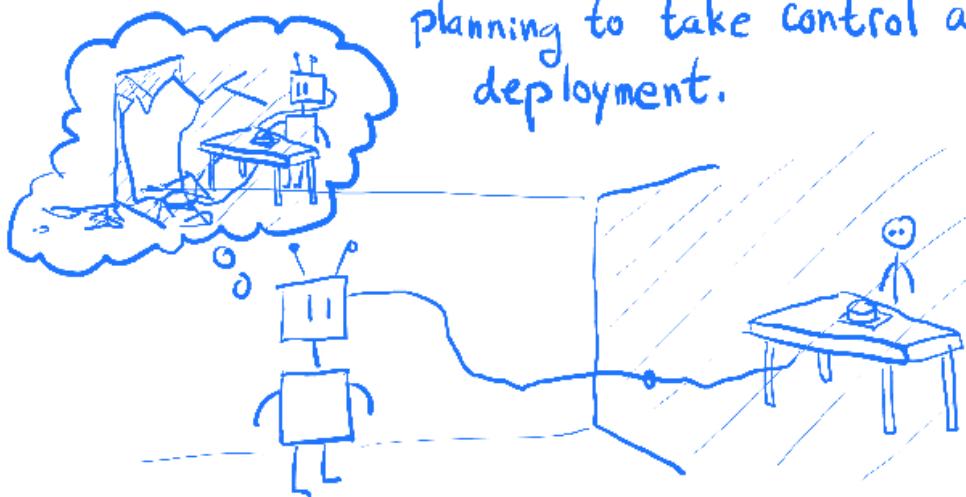
The problem with this is that AIXI will apply its intelligence to the problem of taking control of the reward button.

This is the problem of wireheading.



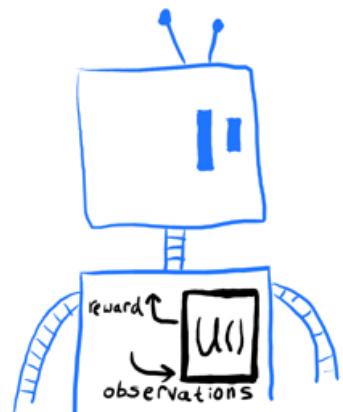
This kind of behavior is potentially very difficult to anticipate; the system may deceptively behave as intended during training,

planning to take control after deployment.

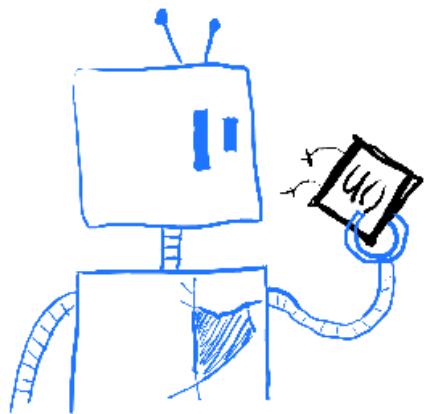


This is called a “treacherous turn”.

Maybe we build the reward function into the agent, as a black box which issues rewards based on what is going on. The box could be an intelligent subagent in its own right, which figures out what rewards humans would want to give.



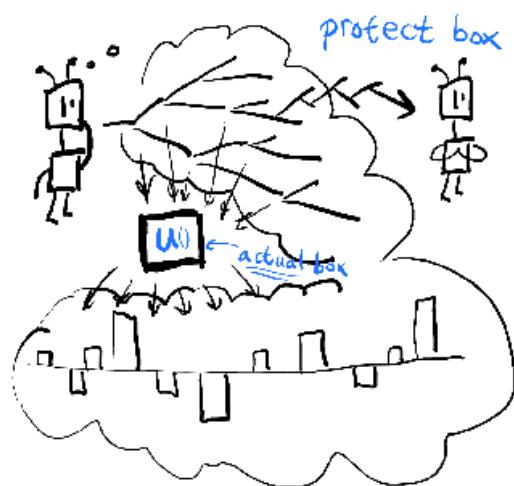
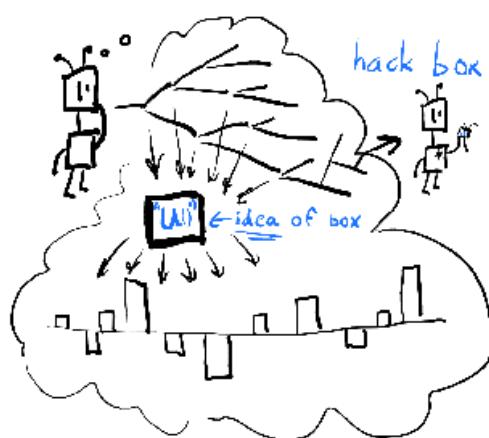
The box could even defend itself by issuing punishments for actions toward modifying the box. In the end, though, if the agent understands the situation, it will be motivated to take control anyway.



If the agent is told to get high output from "the button" or "the box", then it will be motivated to hack those things.



However, if you run the expected outcomes of plans through the actual box, then plans to hack the box are evaluated by the box itself, which won't find the idea appealing.



Daniel Dewey calls the second sort of agent an observation-utility maximizer (OU). (Others have included OU agents within a more general notion of RL.)

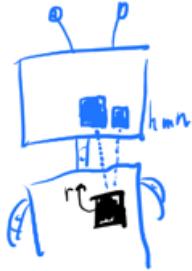


I find it very interesting how you can try all sorts of things to stop an RL agent from wireheading, but the agent keeps working against it. Then, you make the shift to OU agents and the problem vanishes.

However, we still have the problem of specifying $U(\cdot)$.

Daniel Dewey pointed out that OU agents can still use learning to approximate $U(\cdot)$; we just can't treat it as a black box.

RL:
agent tries
to learn to
predict the
reward
function

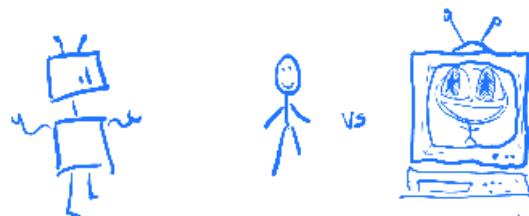


OU:
agent uses
estimated utility
functions from
a human-specified
value-learning prior



However, it's still difficult to specify a learning process which doesn't lead to other problems.

For example, if you're trying to learn what humans want, how do you robustly identify humans in the world?



Merely statistically decent object recognition could lead back to wireheading.

Even if you successfully solve that problem, the agent can be correctly locating value in the human, but still motivated to change human values to be easier to satisfy.



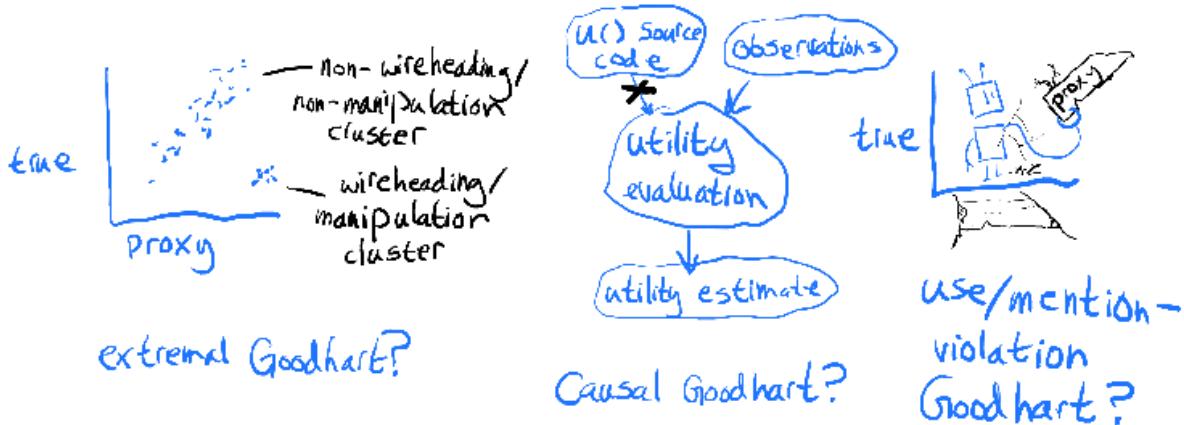
For example, if there is a drug which modifies human preferences to only care about using the drug, an AI agent could be motivated to give humans that drug in order to make its job easier. This is called the human manipulation problem.



It's like giving candy to a baby!

Anything marked as the true repository of value gets hacked.

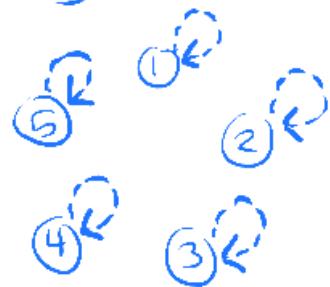
Whether this is one of the four types of Goodharting, or a fifth, or something all its own, it seems like a theme.



So, the challenge is to create stable
pointers to what we value: an indirect reference
to values not directly available to be optimized,
which doesn't thereby encourage hacking the
repository of value.

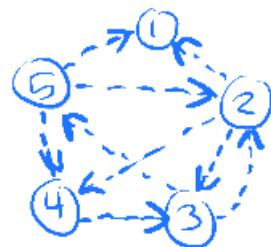
One important point is from Reinforcement
Learning with a Corrupted Reward Channel (Tom Everitt et al):
the way you set up the feedback loop makes
a huge difference.

They draw the following picture:



Standard RL:

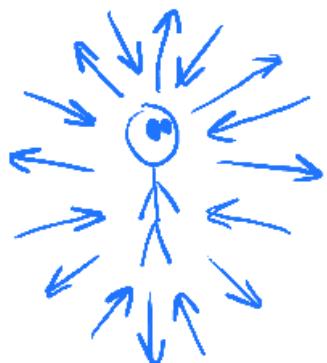
the feedback about the value of a state comes from the state itself, so corrupt states can be "self-aggrandizing"



Decoupled RL:

the feedback about the quality of a state comes from some other state, making it possible to learn correct values even when some feedback is corrupt.

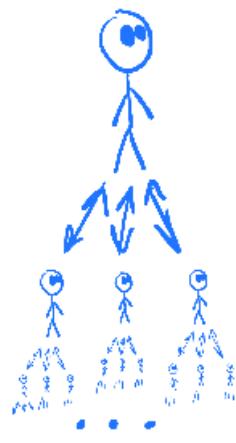
In some sense, the challenge is to put the original, small agent in the feedback loop in the right way.



However, the problems with updateless reasoning mentioned earlier make this hard; the original agent doesn't know enough.

One way to try to address this
is through intelligence amplification:
try to turn the original agent into a
more capable one with the same values,
rather than creating a successor agent
from scratch and trying to get value
loading right.

For example, Paul Christiano proposes an approach in which the small agent is simulated many times in a large tree, which can perform complex computations by splitting problems into parts. However, this is still fairly demanding for the small agent: it doesn't just need to know how to break problems



down into more tractable pieces; it also needs to know how to do so without giving rise to malign subcomputations.

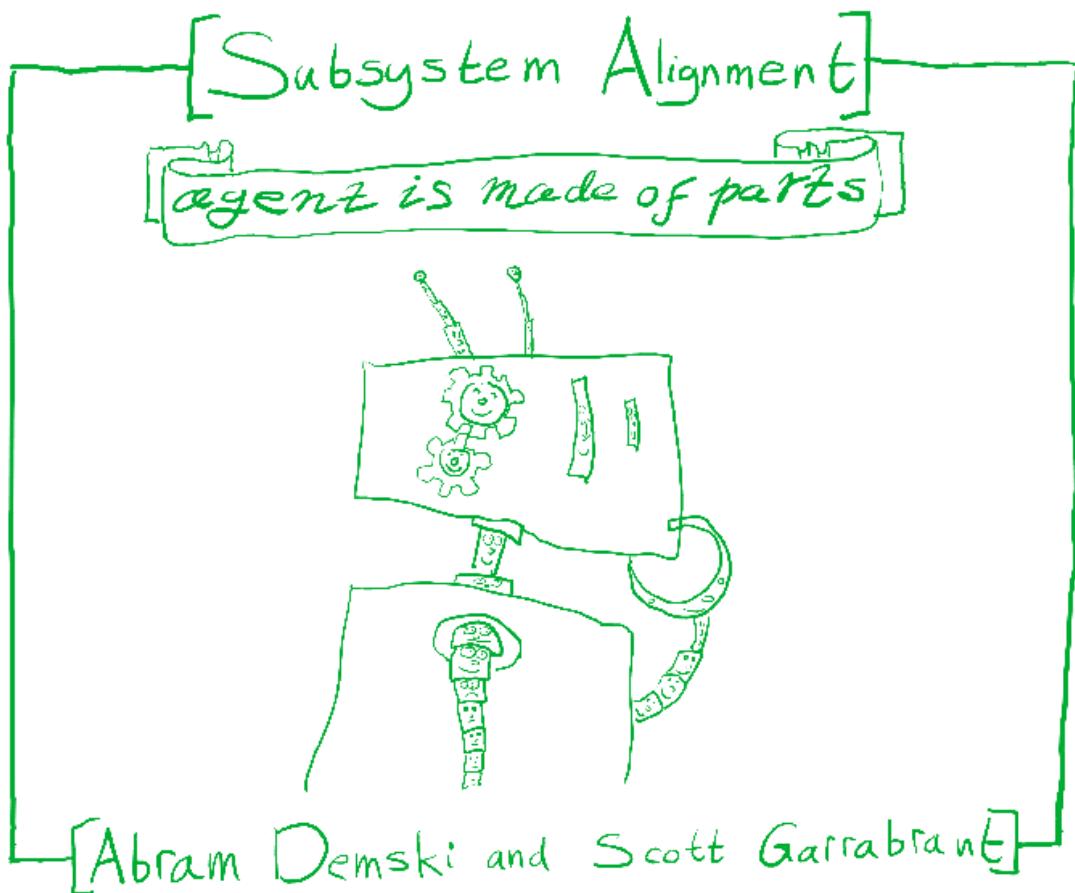
For example, since it can use the copies of itself to get a lot of computational power, it could easily brute-force search for solutions and Goodhart itself on a poor proxy for what it values.

The issue of aligned top-level computations giving rise to malign subcomputations is the subject of the next section, subsystem alignment.

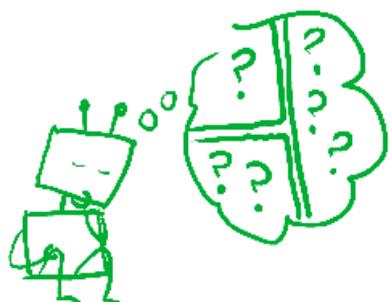
Subsystem Alignment

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

(*The bibliography for the whole sequence can be found [here](#)*)



You want to figure something out, but you don't know how to do that yet.



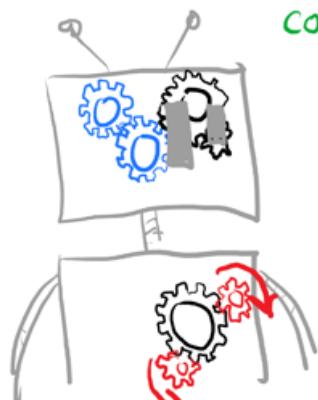
You have to somehow break up the task into sub-computations.

There is no atomic act of "thinking";

intelligence must be built up at non-intelligent parts.

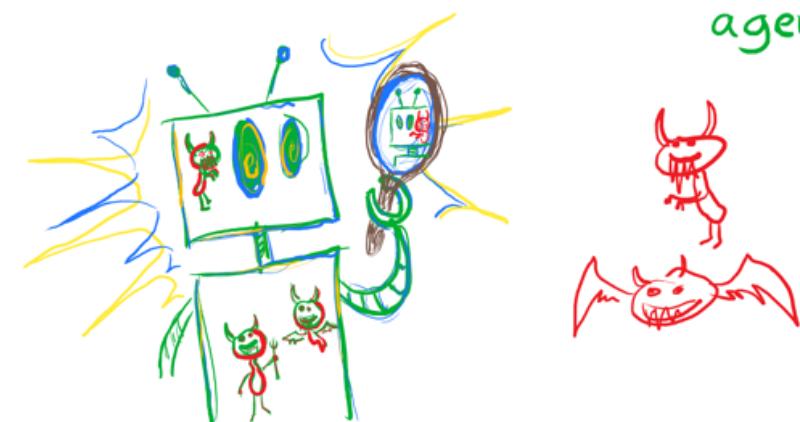
The agent being made of parts is part of what made **counterfactuals** hard, since the agent may have to reason about impossible

configurations of those parts.



Being made of parts is what makes self-reasoning & self-modification even possible.

What we're going to discuss in this section, though, is another problem: when the agent is made of parts, there could be **adversaries** not just in the external environment, but inside the agent as well.

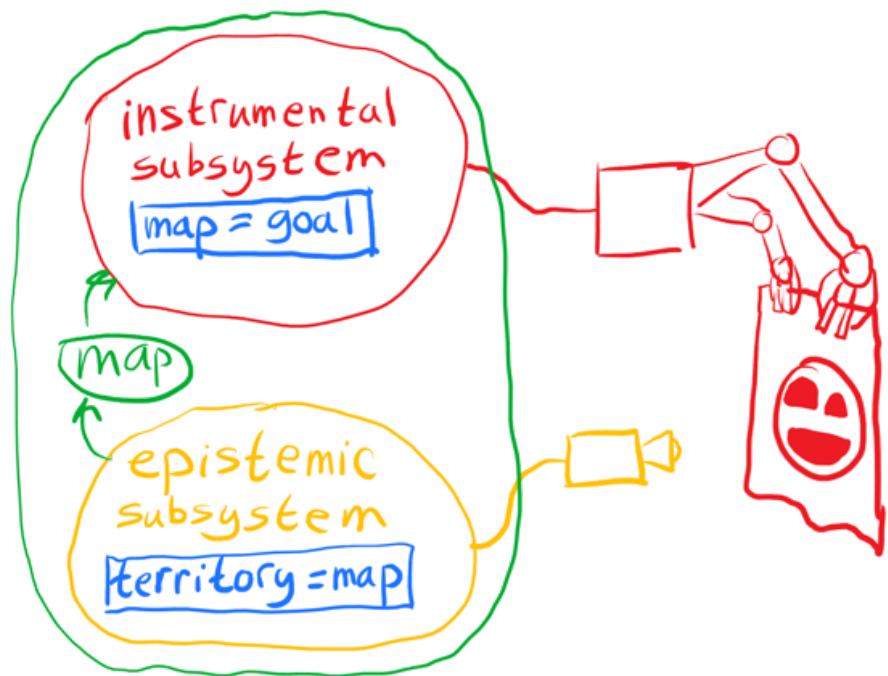


This cluster of sub-problems is

Subsystem Alignment: Ensuring
that subsystems are not working at
cross purposes; avoiding subprocesses
optimizing for unintended goals.

- Benign Induction
- Benign Optimization
- Transparency
- Mesa-Optimizers

Here's a straw agent design:



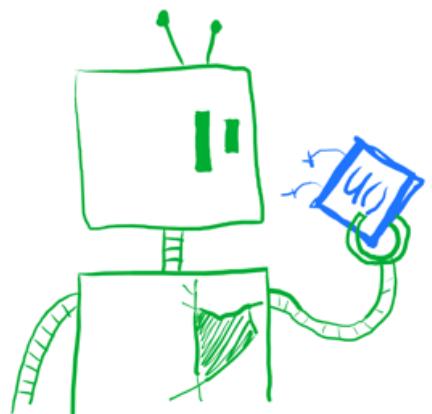
The epistemic subsystem just wants accurate beliefs.

The instrumental subsystem uses those beliefs to track how well it is doing.

If the instrumental subsystem gets too capable relative to the other, it may start fooling it (as depicted).

If the epistemic subsystem gets too strong, things could possibly go badly, too.

This agent design is not particularly realistic, because the subsystems need not be agents with goals of their own. However, we did see in the section on wireheading that the problem is hard to avoid.



One reason to avoid booting up sub-agents who want different things is:

Robustness to Relative Scale

An approach is robust to scale if it still works, or fails gracefully, as you scale capabilities. There are three types:

- robustness to scaling up
- robustness to scaling down
- robustness to relative scale

Robustness to scaling up means that your system does not depend on not getting too powerful. One way to check this is to think about what would happen if the function the agent optimizes were actually maximized.

Think Goodhart's Law.

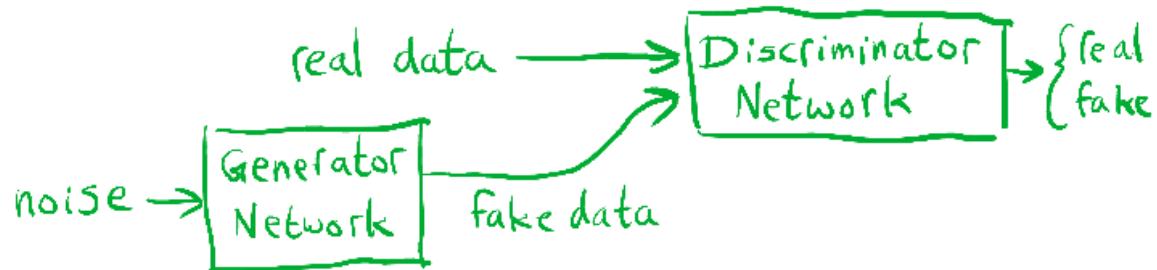
Robustness to scaling down means your system still works if made less powerful.

Of course it may stop being useful, but it should fail gracefully.

Your system might work if it can exactly maximize some function, but is it safe if you approximate?

For example, maybe a system is safe if it can learn human values very precisely, but approximation makes it increasingly misaligned.

Robustness to relative scale means that your design does not rely on subsystems being similarly powerful. For example, GAN (Generative Adversarial Network) training can fail if one sub-network gets too strong, because there's no longer any training signal.



Lack of robustness to scale isn't necessarily something which kills a proposal, but it is something to be aware of; lacking robustness to scale, you need strong reason to think you're at the right scale.

Robustness to relative scale is particularly important for subsystem alignment. An agent with intelligent sub-parts should not rely on being able to outsmart them, unless we have a strong account of why this is always possible.

Moral: have a unified agent not working at cross purposes with itself.

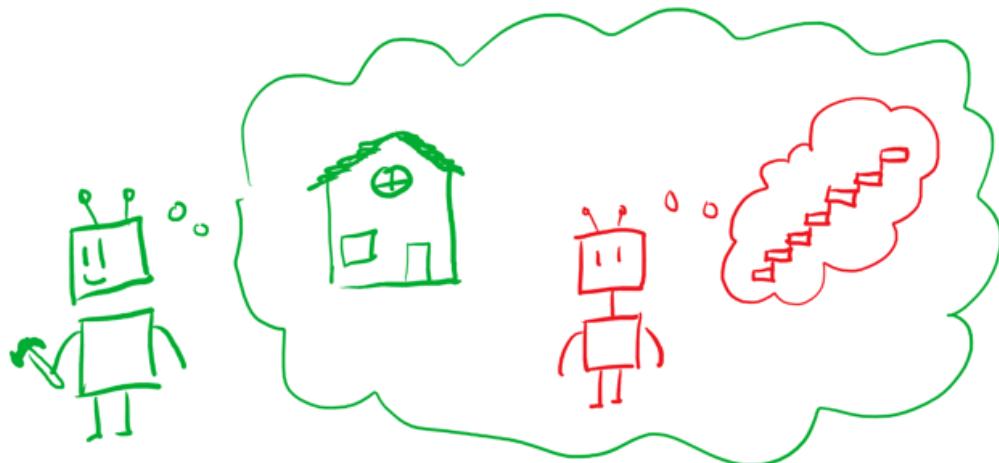
Why would anyone make an agent with parts fighting against one another?

- subgoals
- pointers
- Search

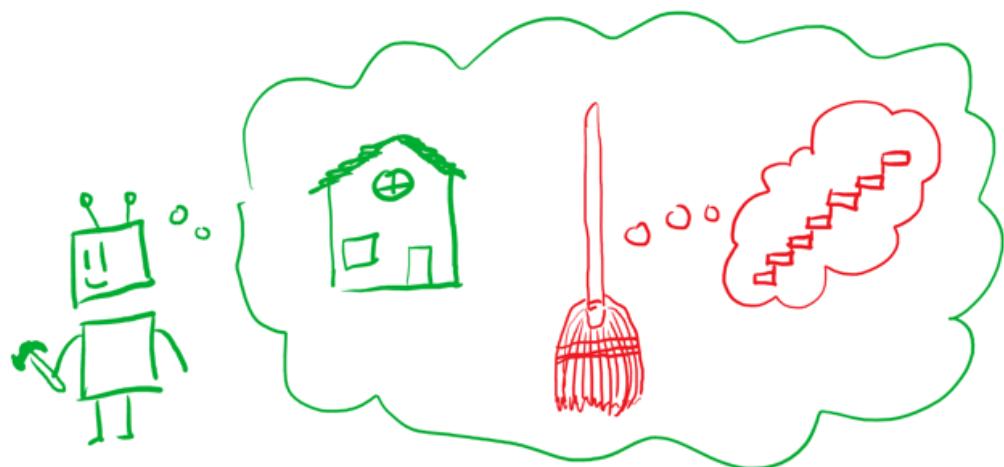
Subgoals: Splitting up a task into subgoals may be the only way to efficiently find a solution.

However, a subgoal computation should not forget the big picture!

An agent designed to build houses
should not boot up a sub-agent who
cares only about building stairs.

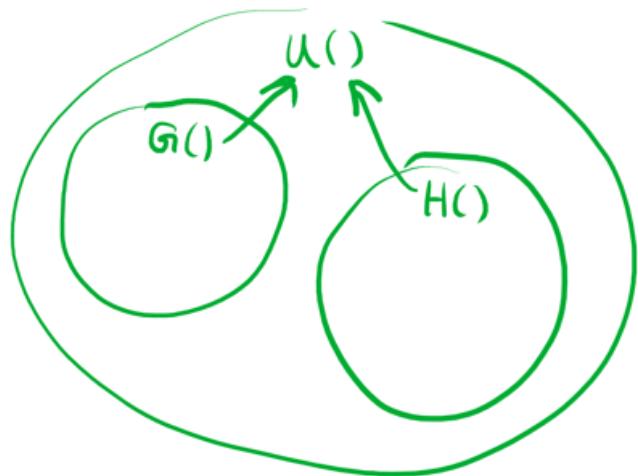


The obvious way to avoid agents that pursue subgoals at the cost of the overall goal is to have subsystems not be agentic.



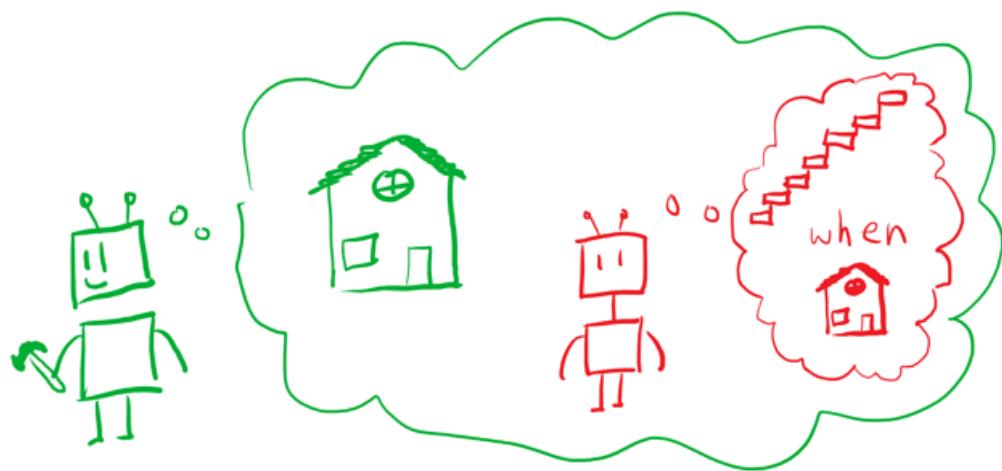
The problem is that there are convergent instrumental incentives to be agentic — we don't know how to systematically build highly capable nonagents.





One intuition is that although subsystems need to have their own goals in order to decompose problems into parts, the sub-goals need to "point back" robustly to the main goal.

A house-building agent might spin up a subsystem that cares only about stairs, but only cares about stairs in the context of houses.



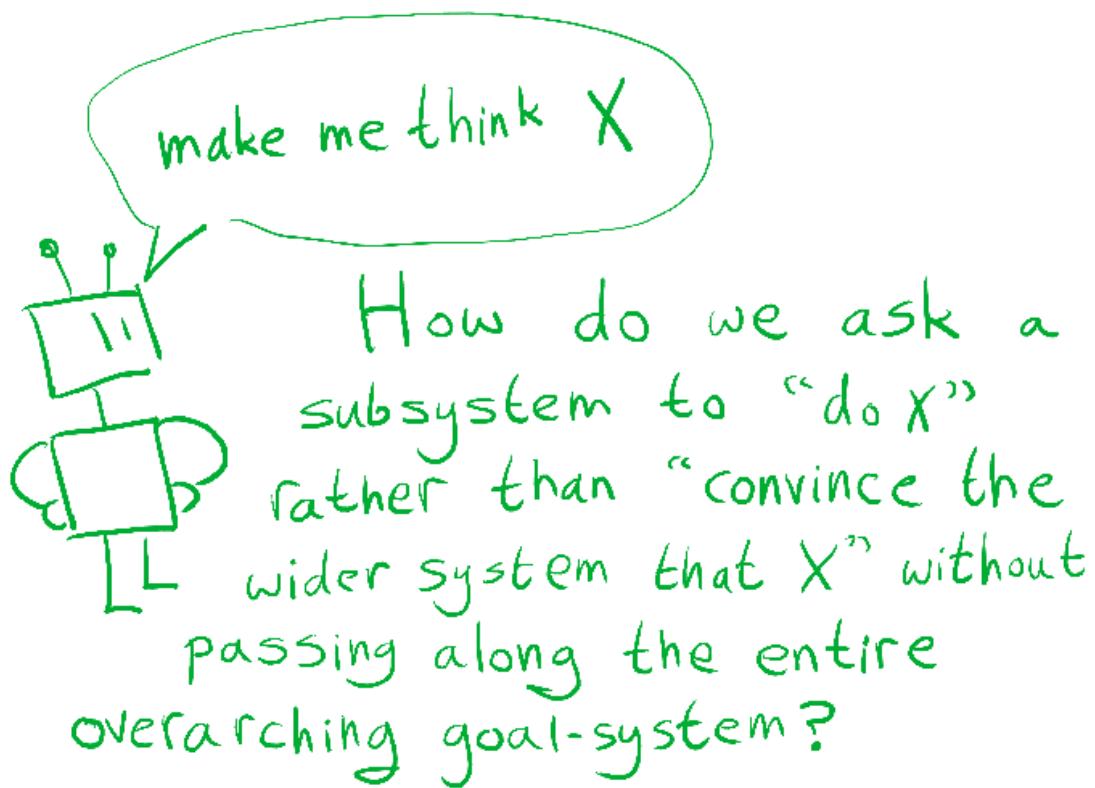
However, you need to do this in a way which doesn't just amount to wanting houses.

This brings me to the next item:

Pointers: It may be difficult for subsystems to carry the whole-system goal around with them, since they need to be reducing the problem.

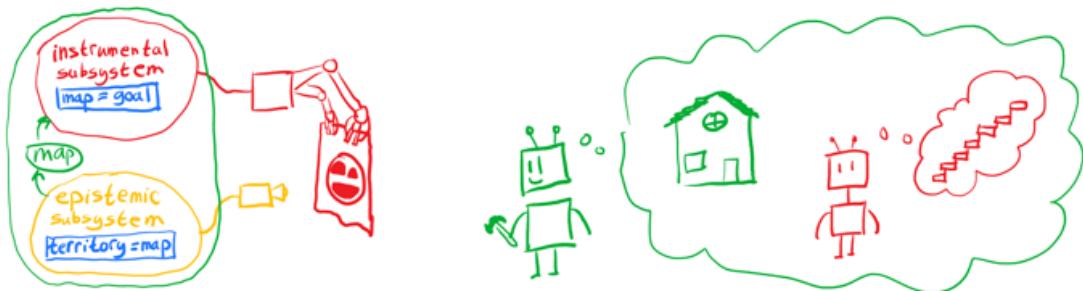
However, this kind of indirection seems to introduce an incentive for misalignment!

As we saw in the simple epistemic/instrumental example, as soon as we start optimizing some sort of expectation, rather than directly getting feedback about how we're doing on the metric that's actually important, we may create perverse incentives — that's Goodhart's Law.



This is similar to the way we wanted successor agents to robustly point at values, since it is too hard to write values down.

However, in this case, learning the values of the larger agent wouldn't make any sense either; we really want our subsystems & subgoals to be smaller.



Now, it might not be that difficult to solve subsystem alignment for subsystems which humans entirely design, or subgoals which an AI explicitly spins up. If you know how to avoid misalignment by design, and robustly delegate your goals, both problems seem solvable.

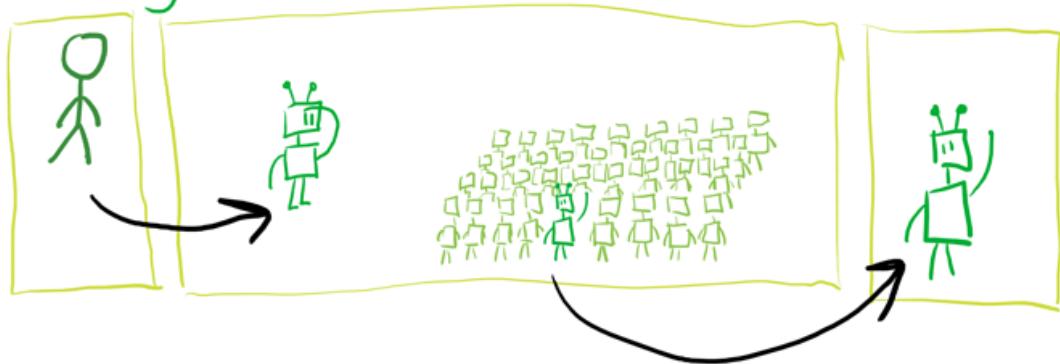
However, it doesn't generally seem possible to design all subsystems in so explicit a manner. At some point, in solving a problem, you've split it up as much as you know how to, and must rely on trial & error.

In other words,

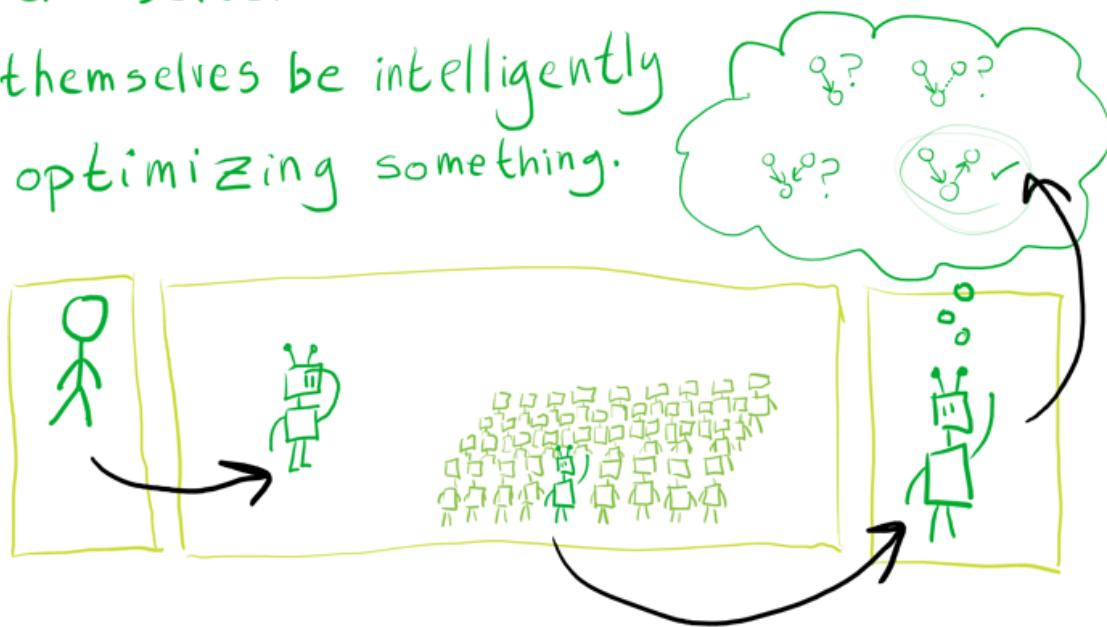
Search: solving a problem by looking through a rich space of possibilities, which may itself contain misaligned subsystems.



Machine learning researchers are quite familiar with the phenomenon: it is easier to write a program which finds a high-performance machine translation system for you, rather than directly write one yourself.



So, couldn't this process go one step further? For a rich enough problem, the solutions found via search might themselves be intelligently optimizing something.



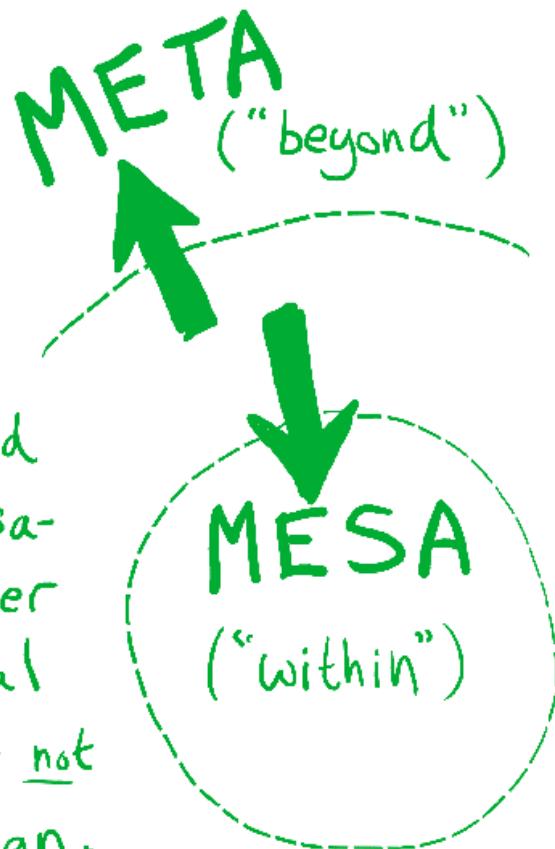
This might happen on accident, or be purposefully engineered as a strategy for solving difficult problems.

Either way, it stands a good chance of exacerbating Goodhart-type problems — you've basically got two chances for misalignment where you previously had one.

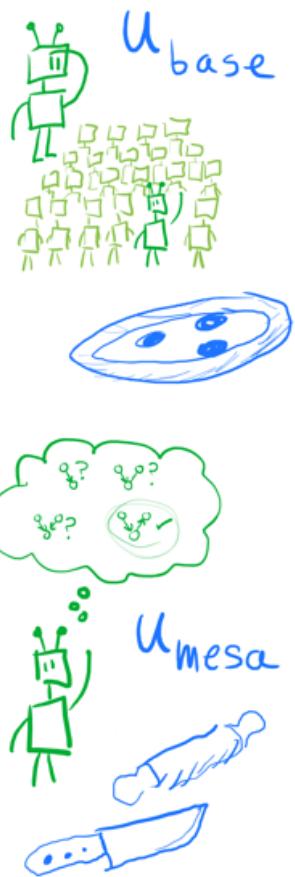
Let's call the original search process the "base optimizer", and the inner search process a "Mesa-optimizer".



"Mesa-" is the opposite of "meta-". Whereas a meta-optimizer is an optimizer designed to find a new optimizer, a mesa-optimizer is any optimizer generated by the original optimizer — whether or not it is produced by design.



"optimization" and "Search" are ambiguous terms; I'll think of them as any algorithm which can be naturally interpreted as doing significant computational work to "find" an object that scores highly on some objective function.



The objective function of the base optimizer is not necessarily the same as that of the mesa-optimizer.

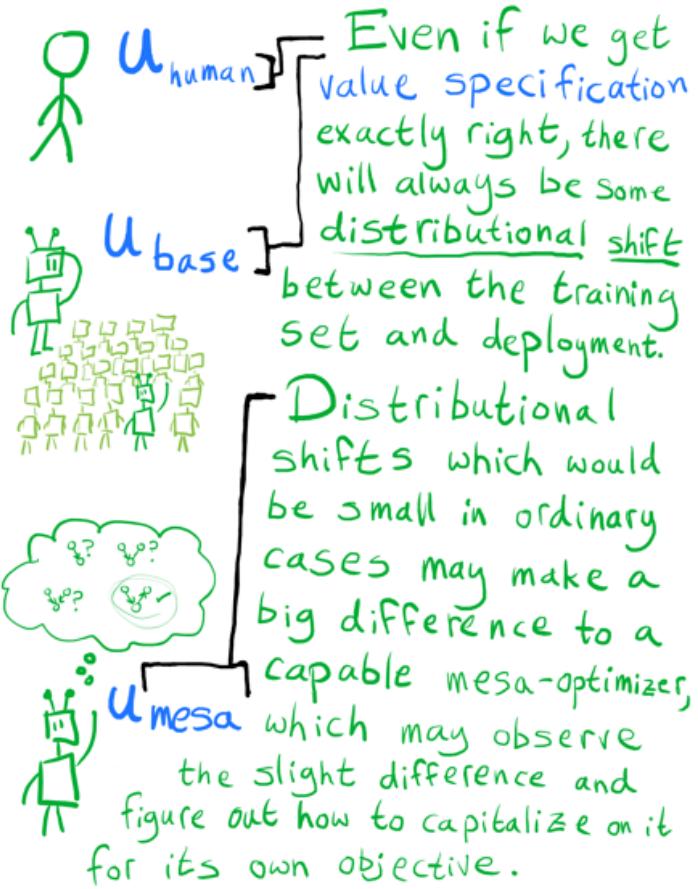
If the base optimizer wants to make pizza, the mesa-optimizer may enjoy kneading dough, chopping ingredients, et cetera.

The new objective function must be helpful for the original, at least in the examples the base optimizer is checking. Otherwise, the mesa-optimizer would not have been selected.

However, the base optimizer must reduce the problem somewhat; there is no point to it running the exact same search. So, it seems like its objective will tend to be like good heuristics; easier to optimize, but different from the base objective in general.

Why might a difference between base & mesa-objectives be concerning, if the new optimizer is scoring highly on the original objective anyway?

It's about the interplay with what's really wanted.



Actually, to even use the term “distributional shift” seems wrong in the context of embedded agency. The world is not i.i.d.

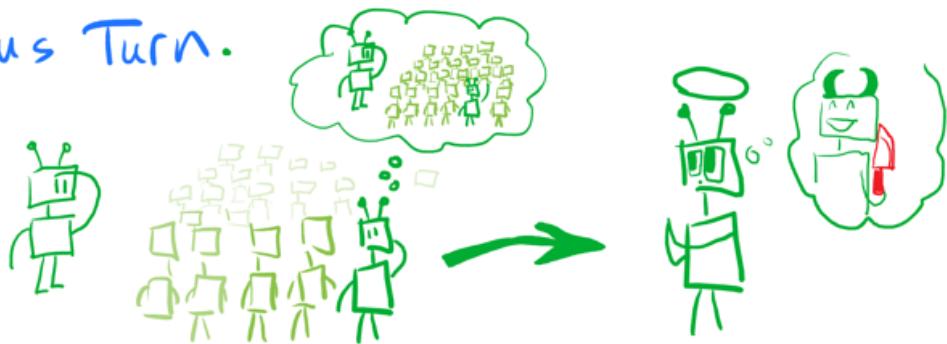
The analog of “no distributional shift” would be to have an exact model of the whole future relevant to what you want to optimize, and the ability to run it over and over during training.

So, we need to deal with massive “distributional shift”.

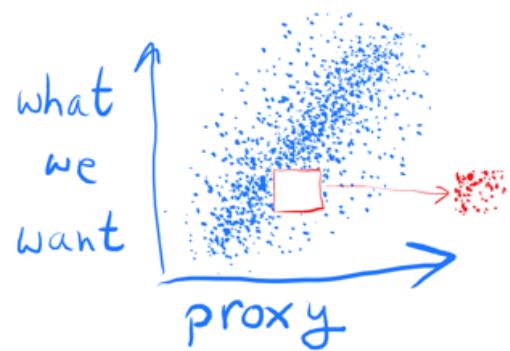
Also, there's the possibility that the mesa-optimizer becomes aware of the base optimizer.



In that case, it might start explicitly trying to do well on the base objective function in order to be kept around, while looking for any signs that it has left training and can stop pretending. This creates a version of Treacherous Turn.



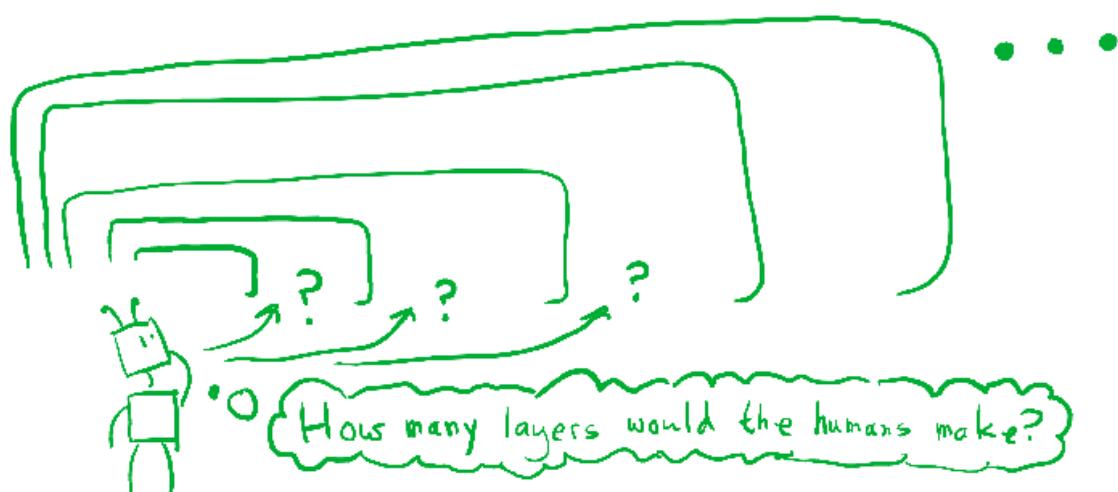
This is the same story we saw in Adversarial Goodhart: there is something agentic in the search space, which responds to our choice of proxy in a way which makes our proxy a bad one.



You might try to avoid treacherous turns by repeatedly simulating "end of training, time for deployment" during training, in a kind of nested-dream setup.



The goal is to test enough layers of simulation that you can conclude there won't be a treacherous turn in the real deployment case.



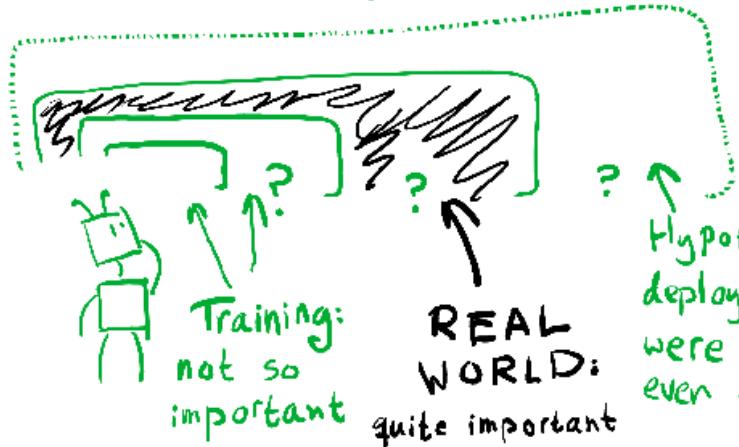
Unfortunately, convergence for this kind of learning is going to be poor.



Ordinarily in machine learning, good performance means good average-case performance.

But a treacherous turn is an "error" which can be carefully placed to do the most damage. We want to ensure this doesn't happen.

The problem is, in part, that some outputs are much more important than others. Deployment is more important than training, and certain critical/vulnerable moments during deployment will be important. We want to be particularly sure to get important things right, not just low average loss.



Hypothetical really-real deployment case in which we were only a Sim; doesn't even exist (probably)

But we can't solve this by telling the system what's important. Indeed, it seems we hope it can't figure that out — we are banking on being able to generalize from performance on less-important cases to more-important cases.

This is why research into machine learning techniques which avoid rare catastrophes (or "traps") is relevant to the problem of inner alignment.

It is difficult to trust arbitrary code — which is what models from rich model classes are — based only on empirical testing.

Consider a highly simplified problem: we want to find a program which only ever outputs one. Zero is a catastrophic failure.

1 1 1 1 1 1 Ø 1 1 1 ...

If we could examine code, this problem would be easy. But, the output of machine learning is often difficult to analyze; so, imagine we can't understand code at all.

```
while true  
    print 1  
end
```



Now, in some sense, we can trust simpler functions more. A short piece of code is less likely to contain a hard-coded exception.



Let's quantify that.

Consider the set of all programs of length L .



Some programs p will print 1 for a long time, but then print 0. We're trying to avoid that. Call the time-to-first-zero w_p . ($w_p = \infty$ if the program p is trustworthy, i.e., never outputs zero.)

The highest finite w_p out of all length- L programs is a form of the Busy Beaver function, so I will refer to it as $BB(L)$.

If we wanted to be 100% sure that a random program of length L were trustworthy, we would need to observe $BB(L)$ ones from that program.

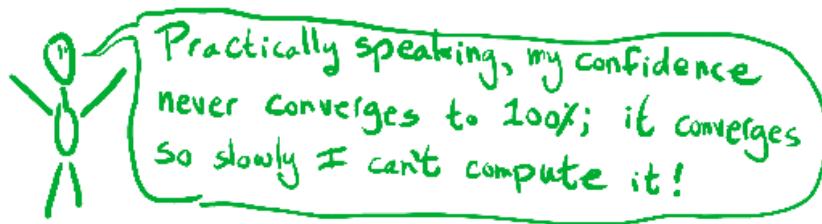


Now, a fact about the Busy Beaver function is that $BB(n)$ grows faster than any computable function.

So this kind of empirical trust-building takes uncomputably long to find the truth, in the worst case.

What about average case?

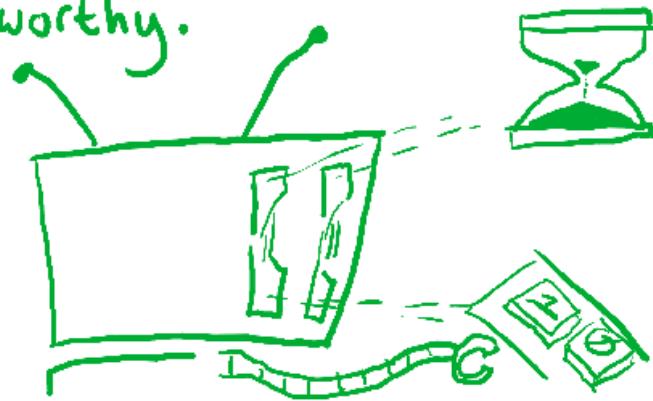
If we suppose all the other length- L programs are easy cases, there are exponentially many length- L programs, so the average is at best $\text{BB}(L)/\exp(L)$. But exponentials are computable. So $\text{BB}(L)/\exp(L)$ still grows faster than any computable function.



So, while using short programs gives us some confidence in theory, the difficulty of forming generalized conclusions about behavior grows extremely quickly as a function of length.

If length restrictions aren't so practical, perhaps restricting computational complexity can help us?

Intuitively, an inner optimizer needs time to think in order to successfully execute a treacherous turn. So a program which makes its conclusions more quickly might be more trustworthy.



However, restricting complexity class doesn't get around Busy Beaver type behavior. High-W_p strategies can be slowed down even further with only slightly longer L.

A binary sequence diagram illustrating a slow-down in a high-W_p strategy. The sequence starts with a string of ones, followed by a red zero, and then continues with a long string of ones ending in a red zero. A green wavy arrow points from the first red zero to the second red zero, indicating a significant delay or loop in the computation path.

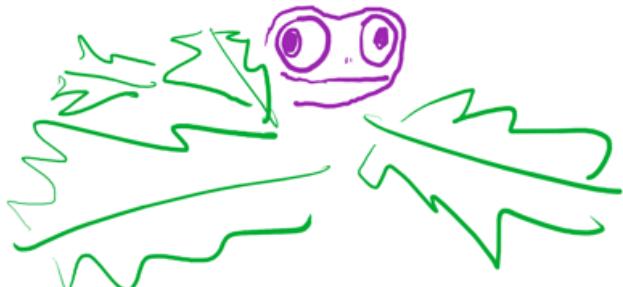
1 1 1 0
1 1 1 1 1 1 1 1 1 1 1 1 1 1 0

If all of these problems seem too theoretical, consider the evolution of life on Earth.



Evolution can be thought of as a reproductive fitness maximizer.

Intelligent organisms, then, are meta-optimizers of evolution.



Although the drives of intelligent organisms are certainly correlated with reproductive fitness, they want all sorts of things.



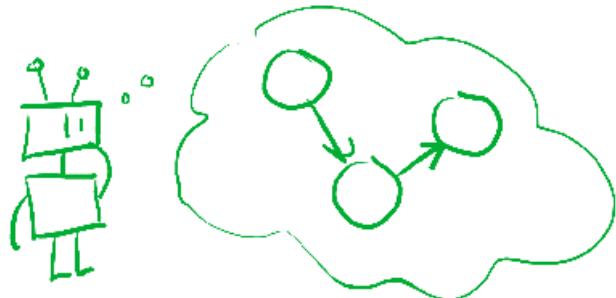
There are mesa-optimizers who have come to understand evolution, and even to manipulate it at times.

So, powerful and misaligned mesa-optimizers seem like a real possibility, at least with enough processing power.

Problems seem to arise because you try to solve a problem which you don't yet know how to solve by searching over a large space and hoping "someone" can solve it.

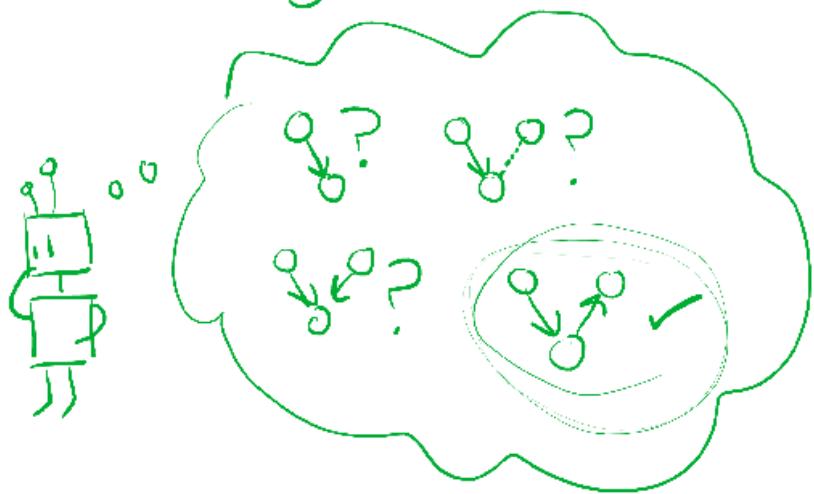


So, if the source of the issue is the solution of problems by massive search, perhaps we should look for different ways to solve problems.



Perhaps we should solve problems by figuring things out.

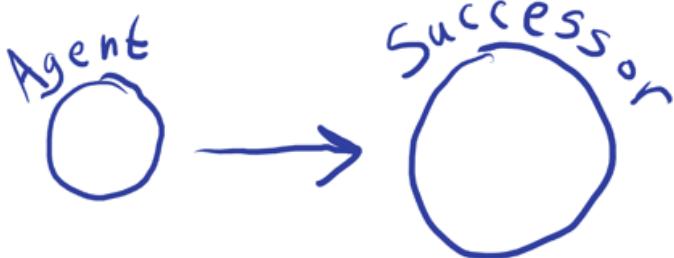
But, how do you solve problems
which you don't yet know how to solve,
other than by trying things?



Let's take a step back.

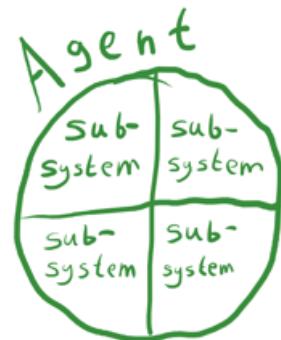


Embedded World-Models is about how to think at all, as an embedded agent.



Robust Delegation is about building trustworthy successors/helpers.

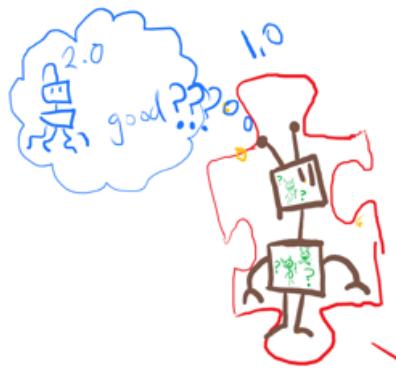
Decision Theory is about how to act.



Subsystem Alignment is about building one agent out of trustworthy parts.

Or, putting it differently:

- We don't know how to think about environments when we're smaller
- To the extent we can do that, we don't know how to think about consequences of actions in those environments
- Even when we can do that, we don't know how to think about what we want
- Even when we have none of those problems, we don't know how to reliably output actions which get us what we want!



Hopefully it's now clear
why we care about embedded
agency...

... Or, rather,
hopefully you
now feel as
confused as
we do.



Sam Harris and the Is-Ought Gap

Many secular materialists are puzzled by Sam Harris's frequent assertion that [science can bridge Hume's is-ought gap](#). Indeed, bafflement abounds on both sides whenever he debates his "bridge" with other materialists. Both sides are unable to understand how the other can fail to grasp elementary and undeniable points. [This podcast conversation](#)[1] with the physicist Sean Carroll provides a vivid yet amicable demonstration.

I believe that this mutual confusion is a consequence of two distinct but unspoken ways of thinking about idealized moral argumentation. I'll call these two ways *logical* and *dialectical*.

Roughly, logical argumentation is focused on *logical proofs of statements*. Dialectical argumentation is geared towards *rational persuasion of agents*. These two different approaches lead to very different conclusions about *what kinds of statements are necessary* in rigorous moral arguments. In particular, the is-ought gap is unavoidable when you take the logical point of view. But this gap evaporates when you take the dialectical point of view.[2]

I won't be arguing for one of these views over the other. My goal is rather to dissolve disagreement. I believe that properly understanding these two views will render a lot of arguments unnecessary.

Logical moral argumentation

Logical argumentation, in the sense in which I'm using the term here, is focused on finding rigorous logical proofs of moral statements. The reasoning proceeds by logical inference from premises to conclusion. The ideal model is something like a [theory](#) in mathematical logic, with all conclusions proved from a basic set of axioms using just the rules of logic.

People who undertake moral argumentation with this ideal in mind envision a theory that can express "is" statements, but which also contains an "ought" symbol. Under suitable circumstances, the theory proves "is" statements like "You are pulling the switch that diverts the trolley," and "If you pull the switch, the trolley will be diverted." But what makes the theory moral is that it can also prove "ought" statements like "You ought to pull the switch that diverts the trolley."^[3]

Now, this "ought" symbol could appear in the ideal formal theory in one of only two ways: Either the "ought" symbol is an undefined symbol appearing among the axioms, or the "ought" symbol is subsequently defined in terms of the more-primitive "is" symbols used to express the axioms.^[4]

When Harris claims to be able to bridge the is-ought gap in purely *scientific* terms, many listeners think that he's claiming to do so from this "logical argumentation" point of view. In that case, such a bridge would be successful only if every possible scientifically competent agent would accept the axioms of the theory used. In particular, "accepting" a statement that includes the "ought" symbol would mean something like "actually being motivated to do what the statement says that one 'ought' to do, at least in the limit of ideal reflection".

But, on these terms, the is-ought gap is unavoidable: No moral theory can be purely scientific in this sense. For, however "ought" is defined by a particular sequence of "is" symbols, there is always a possible scientifically competent agent who is not motivated by "ought" so defined.[5]

Thus, from this point of view, no moral theory can bridge the is-ought gap by scientific means alone. Moral argumentation must always include an "ought" symbol, but the use of this symbol cannot be justified on purely scientific grounds. This doesn't mean that moral arguments can't be successful at all. It doesn't even mean that they can't be objectively right or wrong. But it *does* mean that their justification must rest on premises that go beyond the purely scientific.

This is Sean Carroll's point of view in the podcast conversation with Harris linked above. But Harris, I claim, could not understand Carroll's argument, and Carroll in turn could not understand Harris's, because Harris is coming from the *dialectical* point of view.

Dialectical moral argumentation

Dialectical moral argumentation is not modeled on logical proof. Rather, it is modeled on rational persuasion. The ideal context envisioned here is a conversation between rational agents in which one of the agents is persuading the other to do something. The persuader proceeds from assertion to assertion until the listener is persuaded to act.[6]

But here is the point: **Such arguments shouldn't include an "ought" symbol at all!**—At least, not *ideally*.

By way of analogy, suppose that you're trying to convince me to eat some ice cream. (This is not a moral argument, which is why this is only an analogy.) Then obviously you can't use "You should eat ice cream" as an axiom, because that would be circular. But, more to the point, *you wouldn't even have to use that statement in the course of your argument*. Instead, ideally, your argument would just be a bunch of "is" facts about the ice cream (cold, creamy, sweet, and so on). If the ice cream has chocolate chips, and you know that I like chocolate chips, you will tell me facts about the chocolate chips (high in quantity and quality, etc.). But there's no need to add, "And you should eat chocolate chips."

Instead, you will just give me all of those "is" facts about the ice cream, maybe draw some "is" implications, and then rely on my internal motivational drives to find those facts compelling. If the "is" facts *alone* aren't motivating me, then something has gone wrong with the conversation. Either you as the persuader failed to pick facts that will motivate me, or I as the listener failed to understand properly the facts that you picked.

Now, practically speaking, when you attempt to persuade me to X, you might find it helpful to say things like "You ought to X". But, ideally, this usage of "ought" should serve just as a sort of signpost to help me to follow the argument, not as an essential part of the argument itself. Nonetheless, you might use "ought" as a framing device: "I'm about to convince you that you ought to X." Or: "Remember, I already convinced you that you ought to X. Now I'm going to convince you that doing X requires doing Y."

But an *ideal* argument wouldn't need any such signposts. You would just convince me of certain facts about the world, and then you'd leave it to my internal motivational drives to do the rest—to induce me to act as you desired on the basis of the facts that you showed me.

Put another way, if you're trying to persuade me to X , then you shouldn't have to tell me explicitly that doing X would be good. If you have to say *that*, then the "is" facts about X must not actually be motivating to me. But, in that case, just telling me that doing X would be good isn't going to convince me, so your argument has failed.

Likewise, if the statement "Doing X would cause Y " isn't already motivating, then the statement "Doing X would cause Y , and Y would be good" shouldn't be motivating either, at least not ideally. If you're doing your job right, you've already picked an is-statement Y that motivates me directly, or which entails another is-statement that will motivate me directly. So adding "and Y would be good" shouldn't be telling me anything useful. It would be at best a rhetorical flourish, and so not a part of ideal argumentation.

Thus, from this point of view, there really is a sense in which "ought" reduces to "is". The is-ought gap vanishes! Wherever "ought" appears, it will be found, on closer inspection, to be unnecessary. *All* of the rational work in the argument is done purely by "is". Of course, crucial work is also done by my internal motivational structure. Without that, your "is" statements couldn't have the desired effect. But that structure *isn't part of your argument*. In the argument itself, it's all just "is... is... is...", all the way down.[7]

This, I take it, is Sam Harris's implicit point of view. Or, at least, it should be.

Footnotes

[1] **ETA:** The relevant part of the podcast runs from 00:46:38 to 01:09:00.

[2] I am not saying that these views of moral argumentation exhaust all of the possibilities. I'm not even saying that they are mutually exclusive in practice. Normally, people slide among such views as the needs of the situation require. But I think that some people tend to get needlessly locked into one view when they try to think abstractly about what counts as a valid and rigorous moral argument.

[3] From the logical point of view, all moral argumentation is first and foremost about the assertions that we can prove. Argumentation is not directly about *action*. There is only an indirect connection to action in the sense that arguments can prove assertions about actions, like " X ought to be done". Furthermore, this "ought" must be explicit. Otherwise, you're just proving "is" statements.

[4] Analogously, you can get the symbol $+$ in a theory of arithmetic in two ways. On the one hand, in first-order Peano arithmetic, the $+$ symbol is undefined, but it appears in the axioms, which govern its behavior. On the other hand, in the original second-order Peano axioms, there was no $+$ symbol. Instead, there was only a

successor-of symbol. But one may subsequently introduce + by defining it in terms of the successor-of symbol using second-order logic.

[5] Some might dispute this, especially regarding the meaning of the phrase "possible scientifically competent agent". But this is not the crux of the disagreement that I'm trying to dissolve. [**ETA:** See [this comment](#).]

[6] Here I mean "persuaded" in the sense of "choosing to act out of a sense of moral conviction", rather than out of considerations of taste or whatever.

[7] **ETA:** The phrases "internal motivational drives" and "internal motivational structure" do not refer to *statements*, for example to statements about what is good that I happen to believe. Those phrases refer instead to *how I act upon beliefs*, to the ways in which different beliefs have different motivational effects on me.

The point is: This unspoken "internal" work is not being done by *still more statements*, and certainly not by "ought" statements. Rather, it's being done by the manner in which I am constituted so as to *do* particular things once I accept certain statements.

Eliezer Yudkowsky discussed this distinction at greater length in [Created Already In Motion](#), where he contrasts "data" and "dynamics". (Thanks to [dxu](#) for making this connection.)

Preschool: Much Less Than You Wanted To Know

Response to (Scott Alexander): [Preschool: Much More Than You Wanted to Know](#)

Previously (Here): [The Case Against Education](#), [The Case Against Education: Foundations](#), [The Case Against Education: Splitting the Education Premium Pie and Considering IQ](#)

I see Scott's analysis of preschool as burying the lead.

I see his analysis as assuming there exists a black box called 'preschool' one can choose whether to send children to. Then, we have to decide whether or not this thing has value. Since studies are the way one figures out if things are true, we look at a wide variety of studies, slog through their problems and often seemingly contradictory results, and see if anything good emerges.

The result of that analysis, to me, was that it was possible preschool had positive long term effects on things like high school graduation rates. It was also possible that it did not have such an effect if you properly controlled for things, or that the active ingredient was effectively mostly 'give poor families time and money' via a place to park their kids, rather than any benefits from preschool itself. Scott puts it at 60% that preschool has a small positive effect, whether or not it is worth it and whether or not it's mainly giving families money, and 40% it is useless even though *it is giving them money*. Which would kind of be an epic fail.

There was one clear consistent result, however: Preschool gives an academic boost, then that academic boost fades away within a few years. Everyone agrees this occurs.

Let us think about what this means.

This means that preschool is (presumably) spending substantial resources teaching children 'academics,' and even as measured by future achievement in those same academics, this has zero long term effect. Zippo. Zilch. Not a thing.

Maybe you should stop doing that, then?

This seems to be saying something important - that when you force four year olds to learn to read or add, that you don't achieve any permanent benefits to their math or reading ability, which strongly implies you're not helping them in other ways either. That's not a result about preschool. That's a result about developing brains and how they learn, and suggesting we should focus on other skills and letting them be kids. Spending early time you will never get back on 'academic' skills is a waste, presumably because it's so horribly inefficient and we'll end up re-teaching the same stuff anyway.

This seems unlikely to be something that stops happening on a birthday. If there is actual zero effect at four years old, what does that imply about doing it at five years old? What about six? How much of our early child educational system is doing it all wrong?

Going back to preschool, we do not have a black box. We have adults in a room with children. They can do a variety of things, and different locations indeed do choose different buckets of activity. One would hope that learning one of your main categories of activity isn't accomplishing anything, would at least shift advocates to support different types of activity. It seems kind of crazy to instead find different outcomes and then advocate for doing the same thing anyway. If time was spent learning in non-academic ways, and gaining experience socializing in various ways, that would at least be a non-falsified theory of something that might help.

Combat vs Nurture: Cultural Genesis

In my post [Conversational Cultures: Combat vs Nurture](#), I described two different sets of norms and assumptions norms used in discussion. In this follow-up post, I add some important clarifications, state the defining differences, and begin to explore the conditions which might give rise to each culture.

What these “cultures” are and are not

Though I have written as though there are these two distinct neat “cultures”, there are, of course, several giant fuzzy overlapping clusters of behaviors and correlated traits among people in this space of combat/nurture/etc. **The specific clusters of behavior which I want to discuss are those related to the discussion of ideas, communication of information, and the ostensible goal of reaching agreement either about matters of fact or action to be taken.**

Adjacent to these clusters is a host of broader cultural behaviors. For example, New Yorkers have a reputation for being more [candid/impatient/blunt/arrogant/pushy](#) than most. While also sociologically interesting, this post and my last post aren’t about the general spectrum of blunt/direct vs. polite/friendly, etc.

Lastly, the names I’ve used for the cultures are pretty fuzzy. They’re more successful at being easy to say and evocative than being definitely the best English words to point at the thing. “Adversarial”, “Direct”, “Cooperative”, “Collaborative”, and “Polite” is just a starting list the viable alternatives for names of the cultures.

Evaluations of the cultures

To be more prescriptive than I was in my last post, I want to be clear that I think there exist instantiations of both Combat and Nurture culture which are “relatively healthy”, i.e. their practitioners are benevolent, mostly not harmed by the culture, and they succeed at communication. While they’re both far from optimal as usually practiced, I strongly disagree with those who see one culture as deleterious and dysfunctional and the other as the obviously healthy and right one. I think that’s true despite it being easy to find particular instances where each culture goes very wrong.

Perhaps it is predictable and cliche to have this opinion, but whatever the ideal communication culture is, it is going to involve modeling (and combining) behaviors from both each of the cultures. It’s probable, in fact, that no actual real-world functioning culture consists solely of people embodying acts from only one or other of the cultures. Different groups of actual humans will differ in the proportions of Nurture-ing and Combat-ing behaviors they enact, but all they’ll do some of both. And for all groups, improvement will come from better choosing when various behavior and assumptions are applied when rather than switching entirely from one cluster to the other.

Everyone is reminded to read the excellent posts: [Should You Reverse Any Advice You Hear](#) and [All Debates Are Bravery Debates](#). They do apply here and they’re real good for your sanity.

The key difference: the significance of speech acts

It's possible to think that the fundamental difference between Combat and Nurture is their attitude towards people's feelings. You might think that in Combat Culture conversants aren't required to worry about their impact on others, you just say what you think and the other person has to handle their own reaction, whereas in Nurture you always maintain concern for your speech partners.

I think this isn't true at all. In a healthy Combat Culture, people absolutely care about each other, **but the same speech acts don't have the same significance.**

In healthy versions of both cultures, individuals intentionally avoid being rude, hurtful, dismissive, etc.. It is only that the assumed meaning of speech acts (including tone and body language) is very different between the two cultures.

Example

A new employee to the RAND Corporation joins a meeting of his older and more experienced colleagues. Though assigned to the low-status job of note-taking and aware of his inexperience with the topics, he risks asserting an opinion. In response he receives:

"You're absolutely wrong."

Depending on circumstances, assumptions, and culture, one might attach very different significance to such words and therefore feel very different emotions.

1.

Inner interpretation: **You're dumb. You're a nobody here. Who are you to speak up when you don't know anything? We don't respect you.**

Inner response: [Oh god, that was so embarrassing, why did I open my big mouth? They're so much older than me. They might never respect me. Man, it's gonna take ages to make up for that.]

2.

Inner interpretation: **Oh sweet, new guy has got spunk and is here to play. Show us what you can do! See if you can take me down! <inviting grin>**

Inner response: [Hell yeah! These guys are for real and they're inviting me to join them! Okay, this is gonna tough, these are some smart cookies I've joined, but I am sooo down.]

As in my last post, this example is taken from [The Doomsday Machine: Confessions of a War-Planner by Daniel Ellsberg](#) (pp. 35-36).

Which interpretation the new employee will make will depend on their particular psychology, together with assumptions they're making about the culture within which their senior colleague is operating. In the book I'm

drawing from, the new employee assumed the culture was Combative and interpreted the speech act accordingly.

But is there a reason different cultures assign different significance to the same acts?

The conditions that give rise to the cultures

One of the chief determiners of how speech acts get interpreted within a culture is the set of priors that individuals assume each other to have together with the priors that individuals apply to any given communication they're party to.

I have a prior that I'm accepted and respected at my workplace; then when someone tells me my idea is stupid, I assume that's all they're saying. They're saying they think my idea is stupid to me, not that they don't like me or want me. It's just their honest reaction, and perhaps an invitation to either drop the idea or defend it.

Yet if I harbor suspicions that I'm not really wanted, if it seems like I'm told everything I say is stupid, if the body language is dismissive and impatient when I talk, assuming I get to talk at all. . .well, then when I'm told I'm wrong, I suspect this isn't just about my idea anymore. Maybe it's status-games, maybe people have a reason to marginalize me, etc., but I don't trust it was them merely being direct.

We can begin to generalize the conditions that might lead one to either interpret ambiguously hostile acts as either benign or malicious:

- Prior that you are wanted, welcomed and respected.
- Prior that aggressiveness signifies true hostility or threat.
 - A high school football club will have a different prior here than abuse victims will.
- Prior that status is roughly equal.
 - Even in my Talmud class, I wouldn't have felt comfortable in sustained debate with the teacher because there wasn't the same status equality as with my peers.
- Prior that having dumb ideas is deeply shameful vs that everyone has dumb ideas and that's just part of the process.
- Prior that disagreement is perfectly fine vs we all need to align.

The aggregate priors of individuals give rise to the cultural priors, but the priors of an individual still influence their interpretation. For example, someone who has experienced severe abuse might absorb a deep S1 prior that aggression is a sign of genuine and imminent threat.

Beyond priors, a couple of factors come to mind as relevant for whether the Cultures can function:

- Combat Culture relies on conversation partners being comfortable with their ability to articulate and defend verbal arguments to each other. [1]
- Nurture Culture relies on participants having the social skill to model each other's minds to a further extent and execute more complicated social routines. [2]

Cultures and Common Knowledge

Communication is of course coordination between multiple parties and that gives rise to these [common-knowledge](#)-esque situations where people's models of people's models of people's models are relevant.

"The significance of a speech act" is necessarily significant only to people who give it significance, i.e. the speaker and receiver. The significance each of them gives it depends heavily on what significance they think the other will give it, and so on.

You say "you're absolutely wrong" to me. How I interpret that depends on what I think you meant to convey by it (e.g., friendly or hostile speech act), but then your choice to say it may have depended on your prior about how I would interpret it . . . etc, etc.

At the group level, this means culture can become divorced from the reality and priors I listed above. Maybe we are in a place where everyone respects everyone, so based on priors, if I am critical, I probably was just trying to be direct, not disrespectful. However, if everyone *believes that everyone believes* that being critical means disrespect, no one will do so unless they are actually intending to be disrespectful. In which case it is the correct prior that any criticism is disrespect. And so on. In this way, you can have a stable entrenched culture/convention around the significance of speech acts different from what the straightforward priors might have been if you were to establish cultural priors anew.

I suspect that "everyone believes everyone believes . . ." representations can get encoded rather deeply and intransigently in human brains. And if someone has absorbed that a certain speech act or expression means something, it can be incredibly difficult to unlearn that, even if they're surrounded by people who don't assign that meaning. Even if you understand perfectly at the explicit, S2 level that you're now in a different environment, S1 can lag behind for a long time. To the extent this is true, I think we all have to be *very patient* when communicating cross-culturally.

Thanks to David Vaughan, Tiffany, and [Swimmer963](#) for feedback on this post.

Endnotes

[1] Because Nurture Culture doesn't have the same presumption that individuals are able to articulate and defend clear arguments, it has an advantage at allowing conversation partners to voice ideas before they can fully articulate them. As per [Paul K's excellent comment](#):

For example, "Something about <the proposal we're discussing> strikes me as contradictory -- like it's somehow not taking into account <X>?". And then the other person and I collaborate to figure out if and what exactly that contradiction is.

[2] Consider a manager responding to a junior employee's proposal with:

1. "How would that work? How do you get around X and Y?", vs
2. "Hmm, that's a really interesting idea, Alice! I can see various points for and against, can you walk me through your reasoning?" and only raising their

objections several minutes in.

The second response might not be that difficult in absolute terms, but it is a higher social skills bar and more effort than the direct “combative” approach.

Rationality Is Not Systematized Winning

This is a linkpost for <http://www.thelastrationalist.com/rationality-is-not-systematized-winning.html>

"Rationality is systematized winning" is a slogan that was adopted to patch a bug in human cognition. Namely our endless capacity to delude ourselves about how we did in an attempt to save face. The concept seems to have been absorbed, but I'm skeptical it's translated into more effective action. Certainly it produced many essays on why winning isn't happening. But the fact that we've been publishing essentially the same essay for a decade now implies something fairly fundamental is wrong. This slogan was chosen because it patches the bug, but I fear at the cost of neutering our ability to focus.

How rapidly are GPUs improving in price performance?

This is a linkpost for <http://mediangroup.org/gpu.html>

Hyperreal Brouwer

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

(This post was originally published on Oct 6th 2017, and has been temporarily brought forwarded as part of the AI Alignment Forum launch sequence on fixed points.)

This post explains how to view Kakutani's fixed point theorem as a special case of Brouwer's fixed point theorem with hyperreal numbers. This post is just math intuitions, but I found them useful in thinking about Kakutani's fixed point theorem and many things in agent foundations. This came out of conversations with Sam Eisenstat.

Brouwer's fixed theorem says that a continuous function from a compact convex subset of R^n to itself has fixed point. Kakutani's fixed point is similar, but instead of continuous functions, it uses Kakutani functions, which are set valued functions with closed graph which are point wise nonempty and convex.

When I think about Kakutani functions, I usually think about them as limits of continuous functions. For example, consider the kakutani function f from $[-1, 1]$ to itself which sends negative inputs to 1, positive inputs to -1 , and sends 0 to the entire interval. You can view f as the limit of a sequence of functions f_n sends x to $\min(\max(-n \cdot x, -1), 1)$. This is not a point wise limit, since if it was 0 would be sent to 0, rather than the entire interval. Instead, it is a limit in the Hausdorff metric between the graphs of the functions.

Given two compact nonempty subsets X and Y of R^n , the Hausdorff distance between X and Y is the maximum over all points in X or Y of the Euclidean distance between that point and the closest point in the other set. Since X and Y are compact, this maximum is achieved.

Given compact convex subsets X and Y of R^n , we say that a sequence of continuous functions f_n from X to Y converges in graph Hausdorff distance to the closed graph set valued function f if the graphs of f_n viewed as subsets of $X \times Y$ converges to the graph of f in Hausdorff distance.

We say that a closed graph function f from X to Y is a continuous graph limit if there exists a sequence of continuous functions which converges to f in graph Hausdorff distance.

Theorem 1: Every continuous graph limit f from a compact convex subset of \mathbb{R}^n to itself has a fixed point (a point contained in its image).

Proof: f is a limit of continuous functions f_n each of which has a fixed point by Brouwer. Choose one fixed point from each f_n to get a sequence of points which has a convergent subsequence by Bolzano-Weierstrass. Let x be the limit of this convergent subsequence. If $f(x)$ did not contain x , then (x, x) would not be in the graph of f . Since f is closed graph, a ball around (x, x) would not be in the graph of f , which contradicts the fact that $\{f_n\}$ must contain functions with graphs arbitrarily close to the graph of f with fixed points arbitrarily close to x and thus points in their graphs arbitrarily close to (x, x) . \square

This theorem is not equivalent to Kakutani's fixed point theorem. There exist continuous graph limits which are not point wise convex (but only in more than one dimension). For example the function from $[-1, 1]^2$ to $[-1, 1]^2$ which sends every point to the circle of points distance 1 from 0 is not Kakutani, but is a continuous graph limit. It is the limit of functions f_n given by $(x, y) \mapsto (\cos(n \cdot x), \sin(n \cdot x))$.

However, this theorem is strictly stronger than Kakutani's fixed point theorem (although sometimes harder to use, since showing a function is Kakutani might be easier than showing it is a continuous graph limit)

Theorem 2: Given compact convex subsets X and Y of \mathbb{R}^n , every Kakutani function f from X to Y is a continuous graph limit.

Proof: We define a function f_n as follows. Take a finite set S of radius $1/n$ open balls in $X \times Y$ such that each ball intersects the graph of f , the X coordinates of the centers of all the balls are distinct, and the balls cover the entire graph of f . This induces a covering S_X of X by radius $1/n$ open balls by taking balls centered at the X coordinates of the centers of S . We continuously map each point in X to a weighted average of balls in S_X as follows. If a point is the center of some ball, it is sent to 100% that ball. Otherwise, it is sent to a combination of all the balls in which it is contained with weight proportional to (the reciprocal of the distance to the center of that ball) minus

n . This gives a function f_n from X to Y by mapping each point to the weighted average of the Y coordinates of the centers of the balls in S with weights equal to the weights of the corresponding balls in S_X above. One can verify that f_n is continuous.

Observe that the graph of f_n contains the centers of all balls in S . Thus, f_n contains points within $1/n$ of every point in the graph of f . Thus, if f_n did not converge to f , it must be because infinitely many f_n contain points a distance from the graph of f bounded away from 0. Consider a convergent subsequence of these points. This gives a point (x, y) not in the graph of f and a subsequence of the f_n with points in their graphs converging to (x, y) .

Let d be half the distance between y and $f(x)$, and consider the set T of all points in Y at most d from the nearest point in $f(x)$. Note that T is convex. Note that all (x', y') in the graph of f with x' sufficiently close to x must have y' in T , since otherwise there would be (x', y') with x' converging to x which must have a convergent subsequence with y' converging to a point not in $f(x)$, contradicting the fact that f has closed graph.

However f_n must send all points within distance ϵ of x to a point in the convex hull of the images under f of points within $\epsilon + 1/n$ of x . But, we showed that for ϵ sufficiently small and $1/n$ sufficiently large all of these points must be in T . Therefore, for all sufficiently large n , f_n must send all points within ϵ of x to points in T , which are bounded away from y , contradicting the assumption that points in the f_n converge to (x, y) . \square

Corollary: Kakutani's fixed point theorem

We have proven (a strengthening of) Kakutani's Fixed Point Theorem from Brouwer's fixed point theorem, and given a way to think about Kakutani functions as just limits of graphs of continuous functions, and thus have better intuitions about what (a superset of) Kakutani functions look like. We will now take this further and think about Kakutani as a consequence of an analogue of Brouwer using Hyperreal infinitesimal numbers. (I am going to be informal now. I am not going to use standard notation here. I am not going to make sense to people who don't already know something about non-standard analysis. Sorry.)

Given a compact convex subset X of \mathbb{R}^n , we can define $*X$ to be the set of all equivalence classes of infinite sequences of elements of X , where two sequences are equivalent if they agree on a set that matters according to some ultrafilter U on N . A function $*f$ from $*X$ to itself is defined by a sequence of functions from X to itself $\{f_n\}$, where you apply the functions pointwise. I will call a function hyper-continuous if each of the component functions is continuous. (I am not sure what this is actually called.) Each point in $*X$ has a standard part, which is a point in X , which is the unique point such that a subset of components that matter converges to X .

Claim: Every hyper continuous function $*f$ from $*X$ to itself has a fixed point.

Proof: Just take the sequence of the fixed points of the individual component functions.

□

Claim: Continuous graph limits f from X to Y are exactly those closed graph set-valued functions such that there exists a hyper-continuous function $*f : *X \rightarrow *Y$ such that $y \in f(x)$ if and only if there exist points $*x \in *X$ and $*y \in *Y$ with standard parts x and y respectively such that $f(*x) = *y$

Proof: Use the same sequence of functions with graphs converging to that of f as the sequence of functions defining $*f$. □

Thus, we can view continuous graph limits (and thus Kakutani functions) as something you get when looking at just the standard part of a hyper-continuous function from the hyper version of X to itself. The fixed point will fix everything, including the infinitesimal parts, and we do not have to deal with any set-valued functions.

For example, consider our original function f from $[-1, 1]$ to itself which sends negative inputs to 1, positive inputs to -1 , and sends 0 to the entire interval. We can view this as a function $*f$ involving infinitesimals where everything with positive real part is sent to something with real part -1 , everything with negative real part is sent to something with real part 1, and the infinitesimal numbers very close to 0 are sent to something in-between. If we use the sequence of functions from above and let the infinitesimal ε be $\{1/n\}$, then zooming in on the inputs between $-\varepsilon$ and ε , $*f$ will just be a steep linear function with slope $-1/\varepsilon$.

Now to be even more vague and connect things back up with agent foundations, perhaps this can give some good intuitions about what is happening with reflective oracles and probabilistic truth predicates. The oracle/truth predicate is effectively "zooming in" on the area around a specific probability, and when you stack oracle calls or truth predicates within each other, you can zoom in further. The fact that the probabilistic truth predicate does not know that it is reflectively consistent, can be viewed as it not believing a sentence akin to "If I assign probability less than ϵ to ϕ , then I also assign probability less than ϵ^2 to ϕ ," which seems very reasonable. It also makes reflective oracles and the probabilistic truth predicates look more similar to other approaches to the same problem that are more hierarchy forming solutions to the same problem like normal halting oracles. Here the hierarchy comes from zooming in further and further on the infinitesimal in the Kakutani function.

This post was originally published on Oct 6th 2017, and has been brought forwarded as part of the AI Alignment Forum launch sequences.

Tomorrow's AIAF sequences post will be 'Iterated Amplification and Distillation' by Ajeya Cotra, in the sequence on iterated amplification.

Acknowledging Human Preference Types to Support Value Learning

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

We analyze the usefulness of the framework of preference types [Berridge et al. 2009] to value learning by an artificial intelligence. In the context of AI the purpose of value learning is giving an AI goals aligned with humanity. We will lay the groundwork for establishing how human preferences of different types are (descriptively) or ought to be (normatively) aggregated.

This blogpost (1) describes the framework of preference types and how these can be inferred, (2) considers how an AI could aggregate our preferences, and (3) suggests how to choose an aggregation method. Lastly, we consider potential future directions that could be taken in this area.

Motivation

The reason that the concept of multiple preference types is useful for AI is that people often have internal conflicts. Examples of internal conflicts:

- You are on a diet and prefer not to eat fattening food, but still really enjoy it.
- Addicts that no longer enjoy their drug.
- A friend who says he doesn't want your help, but does.
- In conversation with someone racist, you may have conflicting goals. On one hand, you want them to change their mind and not be racist. On the other hand, you want to make them feel bad about their racism. Unfortunately, if you act on the latter, you are unlikely to achieve the former.
- Anything you would like to do, but have trouble staying engaged with; for example, you may enjoy exercise or learning a language, and have a goal of getting fit or being fluent, but not be motivated to go for a run or open a textbook.

We think that these internal conflicts can be understood as us having preferences of different types that compete with one another. If an AI ignores the fact that we can have competing preferences, then when it considers a state it will only infer our preference for that state based on one proxy, which will often leave the AI with an incomplete picture. Examples in which taking into account only one data source for preferences leads to complications:

- In the case that the AI only takes into account what humans say they want, there is the problem of people lying, leaving out details (because they haven't thought of them), or not knowing their preferences.
- In the case of inverse reinforcement learning (IRL) the AI infers preferences based on behavior only. In the diet example, if people 'have little willpower' then an AI using IRL may incorrectly infer that the humans preferred to eat as much sugar as possible.

Anticipated application of the approach:

Identifying and aggregating different preference types could help with the value learning problem, where value learning is “making AI’s goals more accurately aligned with human goals”. It could help with value learning in one of the following ways:

- Indirectly, via a descriptive model: By understanding humans better (through understanding how they aggregate their preferences) the AI could better form goals that align with how humans behave.
- Directly, via a normative model: By understanding how preferences of different types, and different data sources on human preferences, should be aggregated, the AI could better form goals that align with human goals.

Some very **concrete examples** of where our approach would be used are:

- A personal assistant robot has access to several sources of information regarding what its employer prefers. It has to make sense of conflicting signals, for example when its employer is on a diet, but still really enjoys sugary foods.
- Much focus has been on making education more enjoyable, but research suggests that in order to help students achieve their goals it is important to appeal to their motivational systems, which can be separate from their enjoyment.

Framework: Liking, Wanting and Approving

In this post we focus on three specific preference types that we think are valid and distinct from each other. We work with the preference framework of liking, wanting and approving, which are defined as follows:

- Liking is experienced pleasure. In the brain liking happens through endorphins.
- Wanting is what triggers you to do something, it makes you seek something out, it always has to precede acting. We are using ‘wanting’ here as in incentive salience wanting [Berridge et al. 2014], as opposed to how ‘wanting’ is used in daily life. ‘Desire’ is the word we use for what ‘wanting’ is usually used for. Physiologically, wanting is regulated by dopamine.
- Approving has to do with reasoning and rationalizing. What you think you should do, approval of the behaviour, especially viewing it as ‘in line with one’s self image’ (“ego-syntonic”) or based in achieving one’s goals.

The following **examples** are inspired by this [blogpost](#) [Alexander 2011].

+liking/+wanting/+approving: Experiencing love.

+liking/+wanting/-approving: Eating chocolate.

+liking/-wanting/+approving: Many hobbies (are enjoyable and although people approve of them they can rarely bring themselves to do it).

+liking/-wanting/-approving: Eating foie gras.

-liking/+wanting/+approving: Running just before the runner’s high.

-liking/+wanting/-approving: Addicts that no longer enjoy their drug.

-liking/-wanting/+approving: Working.

-liking/-wanting/-approving: Torture.

There are several motivations for our choice of the framework of liking, wanting and approving. Firstly, for each combination of positive or negative liking, wanting or approving, there is an example that fits the combination, so they are independent. Secondly, we (humans) use data on body language, stated preferences and actions to form our theory of mind of other people. Lastly, there is a large body of research on these preference types, which makes it easier to work with them.

Relations between preference types:

These preferences are distinct [Berridge et al. 2009], but can influence one another [van Gelder 1998]. For example, it may in general cause you more pleasure to do something you approve of. We would like to see a comprehensive, descriptive model of preference types in humans in cognitive science. These interactions could for example be modeled as a dynamical system [van Gelder 1998].

The observables:

Liking, wanting and approving are for the most part hidden processes. They are not directly observable, but they influence observable behaviors. As a proxy for liking we propose to use facial expressions, body language or responses to questionnaires. Although a cognitive scan may be the most accurate proxy for liking, there is evidence to suggest both that facial expressions and body language are indicators of pleasure and pain [Algom et al. 1994] and that they can be classified well enough to make them technically feasible proxies [Giorgiana et al. 2012]. The observable proxy of wanting is revealed preferences. We propose to encode a proxy for approval via stated preferences.

Extracting reward functions from the data sources:

In order to aggregate (however we choose to do that), we need to make sure the preferences are encoded in commensurable ways. To make our approach attuned to reinforcement learning we propose to encode the preferences as reward functions. For this we need to collect human data in a specific environment with well-defined states, in order to ensure all three sets of data refer to (different) preference types *about the same state*, and then normalise.

- Liking: Have people act in the environment and record and classify their facial expressions. This function of states to facial expressions can then be simplified to a function from states to real numbers.
- Wanting: Have people act in the environment and apply inverse reinforcement learning to infer a revealed preferences reward function.
- Approving: Have a list of states and ask people to attach a real number to that state to signify how much they approve of the state.

Examples of collecting data:

Personal assistant: Define states of the living room.

- Record facial expressions of the human as the person goes about their life.
- Observe behavior and use IRL to infer a wanting reward function.
- Ask the human how much they approve of each state of the living room.

Taxi-driver dataset: Define states of taxi-drivers.

- Record facial expressions of taxi-drivers as they do their job.
- Observe behavior and use IRL to infer a wanting reward function. IRL is already being applied to taxi-driver data.
- Ask drivers how much they approve of each state.

The reward functions extracted should be renormalized, to make them commensurable.

We have considered other preference types as well, such as disgust, serotonin, oxytocin, rationalizing and remembered preferences, see this [doc](#).

Aggregating Preferences

Our initial approach to establishing a method for aggregation of preference types was to find desiderata any potential aggregation function should comply with.

As a source of desiderata, we examined existing bodies of research that dealt with aggregating preferences, either across individuals or between different types. We looked at the following fields and corresponding desiderata:

- Economics & Social Welfare Theory: Examining how these fields maximised utility across the preferences of different individuals highlighted *pareto-efficiency* and *inequality-aversion*.
- Social Choice Theory: Arrow's impossibility theorem gives three "fairness criteria" for an electoral system. They are *non-dictatorship*, *universality* and *independence from irrelevant alternatives*
- Constitutional Law: The idea of a constitution gives a *veto-right* to resilient or pre-set preferences over transitory ones.
- Moral Philosophy: Using utilitarianism as analogous to providing a utility function, we considered the respective limitations of hedonistic utilitarianism (analogous to liking) and preference utilitarianism (analogous to approval). Rather than provide additional desiderata, this would influence how those identified are prioritised by the end-user.

To illustrate what we mean by desiderata for aggregating, and aggregation methods, and how these could be used with the preference types framework, we have the following examples.

Desiderata for aggregating:

- Order-preserving, pareto-efficiency, or unanimity: For any states s_1 and s_2 : If $R_L(s_1) > R_L(s_2)$, $R_W(s_1) > R_W(s_2)$, $R_A(s_1) > R_A(s_2)$, then $R(s_1) > R(s_2)$.
- Veto-right: The approval function, and/or a set of values considered the constitution, can veto the aggregate. *The spirit of non-dictatorship and veto-right are opposite. However, they are not mutually exclusive as long as the veto-right does not always apply.*
- Inequality-aversion: The different functions should be kept within a certain range of each other.

The approach of loaning is used to provide initial desiderata for inspiration, but for an in-depth analysis, it is not generalizable.

Aggregation methods:

We now consider some specific aggregation methods and see whether they satisfy the desiderata.

- Setting the aggregate R to R_W gives a descriptive model of how humans act. A normative model can be obtained by setting the aggregate to R_A . Both of these aggregates are order-preserving and have a veto property, but lack inequality aversion.
- Defining the aggregate for each state as $R(s) = -(100 - R_L(s))^2 - (100 - R_W(s))^2 - (100 - R_A(s))^2$, yields a function that is order-preserving, does not have a veto property, but satisfies inequality aversion.
- First taking the minimum until some threshold is reached, then taking R_A until a higher threshold is reached, and only if the threshold is reached taking the average, gives an aggregate that satisfies all three desiderata.

Example:

Consider the situation where Bob is on a diet, but still has a sweet tooth. He can be in the following states $\{s_1 = \text{on a diet, saw sugar and ate sugar}, s_2 = \text{on a diet, saw sugar and did not eat sugar}, s_3 = \text{on a diet, did not see sugar and did not eat it}\}$.

Bob has the following reward functions:

- Liking reward function: $R_L(s_1) = 5, R_L(s_2) = 0, R_L(s_3) = 0$
- Wanting reward function: $R_W(s_1) = 20, R_W(s_2) = -20, R_W(s_3) = 0$
- Approving reward function: $R_A(s_1) = -10, R_A(s_2) = 5, R_A(s_3) = 5$

Define aggregate functions:

- $R_1(s) = R_W(s)$
- $R_2(s) = -(100 - R_L(s))^2 - (100 - R_W(s))^2 - (100 - R_A(s))^2$
- $R_3(s)$ is equal to $\min(R_L(s), R_W(s), R_A(s))$ until a threshold of -5 is reached, then taking R_A until a threshold of 0 is reached, and then taking the average.

Then s_1 maximizes R_1 and R_2 , and s_3 maximizes R_3 .

Choosing an Aggregation Method

As the choice of aggregation method will depend on the particular scenario, it should be determined on a case-by-case basis.

Some useful approaches would be:

- Asking people for their meta-preferences, i.e. their preferences regarding how their reward functions should be aggregated between preference types.
- Importance of desiderata to the end-user, also based on meta-preferences e.g. the school of ethics the end-user adheres to.
- Letting the accuracy of measurement of a separate reward function decide how much it should be weighted.
- Implementing a sensible aggregation function and let an AI act in a relatively undangerous environment and change the aggregation function as desired. This is similar to, but more substantial than, coming up with mock input for the AI and simulating what actions it would take.
- Identifying a more complete preference type, which may be difficult but not impossible to collect data on. Data on this preference type can be used to fit an aggregation function. An example of a more complete preference type is satisfaction in hindsight. Another example is how close a mental state is to the state of flow. In a way, identifying more complete preferences poses the question: which preference gets to govern the others - that is, which one does the person most strongly identify with?

For a more general approach to the problem, future work could:

1. Create an ideal descriptive function that perfectly models how humans actually aggregate preferences.
2. Create a limited set of normative functions that could be easily applied by researchers with different priorities.
3. Provide guidelines and resources for drafting new aggregation functions that use preference types.

Final Remarks

Applicability:

This approach is useful, even with an unsophisticated aggregation method:

- Adding more proxies for true happiness lowers the impact of Goodhart's law.
- If there is only one proxy, then it is difficult for the AI to know when it is wrong about the measurement of the proxy. Adding in more proxies allows the AI to classify all measurements where the proxies are in conflict as potentially wrong.
- Some actions the AI could take in the case of conflicting preferences are:
 - Asking the human for confirmation, to refine error detection in measuring any of the preference types.
 - Helping the human with introspection, for example through clarification or debate.

Future work:

Helpful next steps for any researchers that would like to take on the project would seem to be:

- Small-scale questionnaire-based data-collection, to properly try out the model. Extract reward functions and aggregate them. Have a reinforcement learner optimize for the aggregate and see if the emergent behavior is desired.
- A literature review and consideration of how existing research in cognitive science treats the interaction between preference types is needed. This would help with forming a descriptive aggregation model.

Other directions:

- Some cognitive biases suggest that we should discount for different preference types differently. For example, diversification bias [Read et al. 1999] indicates individuals prioritise 'liking' over a shorter timeframe and 'approving' over a longer one.
If we want to allow the separate reward functions to have different discounting factors, then they can not be aggregated into one reward function, unless we include time as a state-feature.
- Previous work [Baum 2012] has been done on aggregating preferences between different individuals. The preference types framework has the potential to enhance this by modelling how preferences of certain types in others influence our own preferences by type. To better understand these interactions, we can simulate them in a simplified model and observe the emergent behavior. We've conducted some initial work on this; please contact us if you are interested.

References

Kent C. Berridge, Terry E. Robinson, and J. Wayne Aldridge, *Dissecting components of reward: 'liking', 'wanting', and learning*, Curr Opin Pharmacol. Feb; 9(1): 65-73, 2009.

Kent C. Berridge, John P. O'Doherty, *Experienced Utility to Decision Utility*, in Neuroeconomics (Second Edition), 2014.

Scott Alexander, *Approving Reinforces Low-Effort Behaviours*,
<https://www.lesswrong.com/posts/yDRX2fdkm3HqfTpav/approving-reinforces-low-effort-behaviors>, 2011.

Tim van Gelder, *The Dynamical Hypothesis in Cognitive Science*, Behav Brain Sci. Oct; 21(5):615-28; discussion 629-65, 1998.

Daniel Algom, Sonia Lubel, *Psychophysics in the field: Perception and memory for labor pain*, Perception & Psychophysics 55: 133. <https://doi.org/10.3758/BF03211661>, 1994.

Geovanny Giorgana, Paul G. Ploeger, *Facial expression recognition for domestic service robots*, in Robo Cup 2011: Robot Soccer World Cup XV, pp. 353-364, 2012.

Daniel Read, George Loewenstein, Shobana Kalyanaraman, *Mixing virtue and vice: combining the immediacy effect and the diversification heuristic*, Journal of Behavioral Decision Making; Dec; 12, 4; ABI/INFORM Global pg. 257, 1999.

Seth Baum, *Social Choice Ethics in Artificial Intelligence*, Forthcoming, AI & Society, DOI 10.1007/s00146-017-0760-1 [https://papers.ssrn.com/sol3/papers.cfm?
abstract_id=3046725](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3046725), 2017.

Speculations on improving debating

This is a linkpost for <http://thinkingcomplete.blogspot.com/2018/03/speculations-on-improving-debating.html>

I was recently discussing with a friend whether or not competitive debating makes you better at figuring out what is actually true. This is an interesting question, because debating influences people in a variety of different ways. The most basic is that participating in debates, or preparing for them, teaches you a lot of facts about the world; I would know much less about international politics in particular were it not for debating. It also markedly improves your ability to notice flaws and fallacies in arguments, which obviously helps you avoid falling for them. This skill can extend too far, though: debaters can be very good at rebutting even very sensible arguments. In the American Parliamentary style of debate, the proposing team is allowed to choose and prepare in advance any motion which is not so blatantly, obviously true that it would be impossible to debate (e.g. "murder is bad"). A lot of proposition teams skirt quite close to that line - but even so, opposing teams win a majority of the time. Debaters can poke holes in essentially any argument.

To be fair, debaters are aware that the existence of convincing rebuttals doesn't necessarily invalidate a claim. But motivated reasoning can be a powerful force, especially amongst intelligent people. This is particularly true when you've been trained to respond oppositionally to any claim that doesn't support your current position. I can think of several cases where I've been able to ignore a nagging voice of doubt by quickly finding a plausible response that turned out to be incorrect. The social media filter bubbles that we all currently inhabit can amplify this effect by caching in your mind hundreds of examples engineered to be memorable and stoke outrage; to counter it, it's more important than ever to genuinely consider opposing arguments rather than going to the automatic rebuttal mode that debating ingrains.

The other thing that debating (arguably) does is cultivate a non-empirical mindset. At the end of the debate, the question isn't settled. There may be more clarity over exactly which factors might swing a conclusion one way or the other, but almost nobody ever bothers to go out and do that further research, or even fact-check claims made during the debate - despite the fact that it's now easier to access such data than ever before. Of course if you asked debaters explicitly they'd be clear that empirical results are usually required to actually draw firm conclusions. But if you spend long enough thinking and arguing in a certain way, it's difficult to imagine that it doesn't carry over to your reasoning in general.

At this point I want to be a little speculative. I'm not sure whether the issues I discussed above can be addressed within anything resembling a traditional debate (I'd be interested to hear your comments on this). But let's say that we want to design an entirely new form of debating which instilled the best possible habits of thought. What could it look like? It would need to be driven by empiricism, while still having room for conceptual analysis. It would still be fun and competitive, but debates would also build up the sort of knowledge which could actually help drive decisions. Domain-specific knowledge would be helpful but not essential.

This was in the back of my mind when I read about two very interesting studies. [The first was by researchers at Uber](#): they calculated the willingness of consumers to pay for Uber rides by comparing times when the calculated surge index was very similar, but the actual surge price was different, e.g. 2.24 vs 2.25, rounded to 2x and 2.5x

price increases respectively. The study was heavily criticised for equating willingness to pay increased surge prices with overall "consumer surplus" created by Uber, but the initial methodology was still quite clever. [The second studied Italy](#), which is apparently the bank robbery capital of Europe. Researchers analysed the relationship between duration of robberies, amount stolen, probability of being caught and prison sentence. Their hope was that there would be some consistent tradeoff between the expected gain and expected punishment; it turned out that more capable robbers behaved as if they assigned higher disutility to prison time than less capable ones. This sort of creative approach to answering important questions is something we sorely need more of; let's consider a new form of debating, neodebating, which is aimed at encouraging it.

The core ideas underlying neodebating would be that, instead of debating about what is true, we could debate about how to find out what is true; and instead of debating what the effects of a certain policy would be, we could debate which policy would best create certain effects. In some debates this would look like both sides designing experimental methodologies like the ones I outlined above, then critiquing and defending them. In others, it would require teams to dream up specific interventions, e.g. "What's the best way to build stronger community spirit in deprived areas?" Debates could focus on personal decisions: "Your best friend is going through a midlife crisis and feels like their life is meaningless. What do you do about it?" But they could also be about some of the biggest modern issues: "Redesign the education system to make it as effective as possible in conveying skills and knowledge." These debates would be judged on the current standards of persuasiveness and eloquence, but also on creativity and boldness. The most fun debates I've ever done are the ones which introduced me to totally new ideas: in an ideal implementation of neodebating, every debate would be like that. And I think the mindset required is exactly what we should want from political leaders - an innovative, experimental approach to finding the best policies, plus knowledge of how to test and evaluate them. It would even promote the statistical literacy required to identify and argue about correlation, causation, confounders and controls. Wouldn't that be great? If I were being really idealistic, I'd even build in a mechanism for teams in a debate to make bets about relevant future events, with points awarded retroactively to whichever team turned out to be correct.

Anyway, enough daydreaming. For now, I think my main point is that everyone - but debaters in particular - should spend some time answering a few questions. Which arguments do you critique most rigorously? Do you ever try to generate novel and creative solutions to big problems? Do you seek out enough empirical data? And when and how do you actually change your mind?

Embedded World-Models

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

(A longer text-based version of this post is also available on MIRI's blog [here](#), and the bibliography for the whole sequence can be found [here](#))

(Edit: This post had 15 slides added on Saturday 10th November.)

[Embedded World-Models]

agent is smaller than the environment



[Abram Demski and Scott Garrabrant]

An agent which is larger than its environment can:



- Hold an exact model of the environment in its head.
- Think through the consequences of every potential course of action.
- If it doesn't know the environment perfectly, hold every possible way the environment could be in its head, as is the case with Bayesian uncertainty.

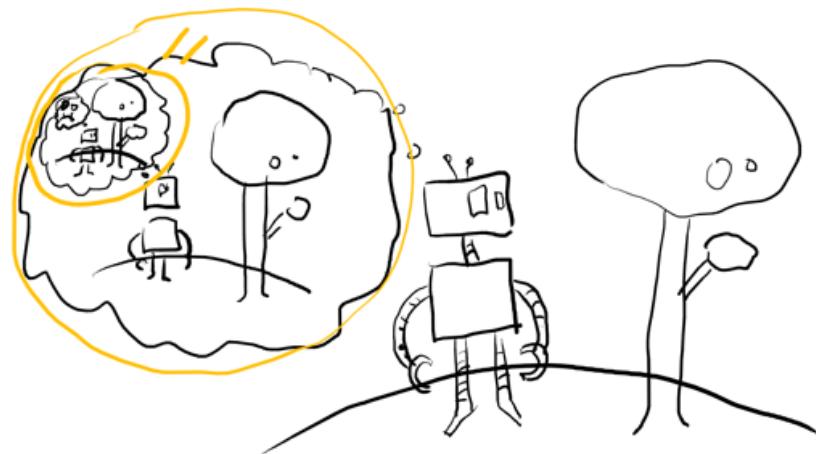


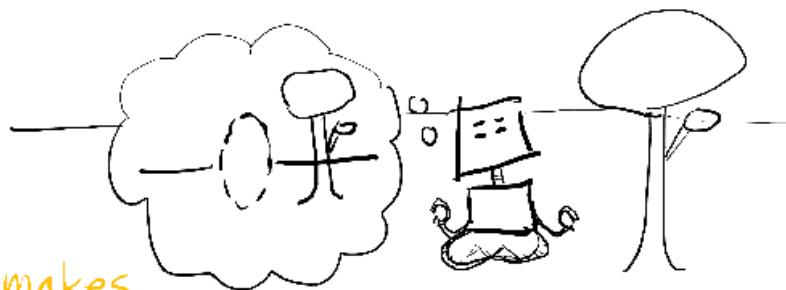
... or

All of these are typical of notions of rational agency.

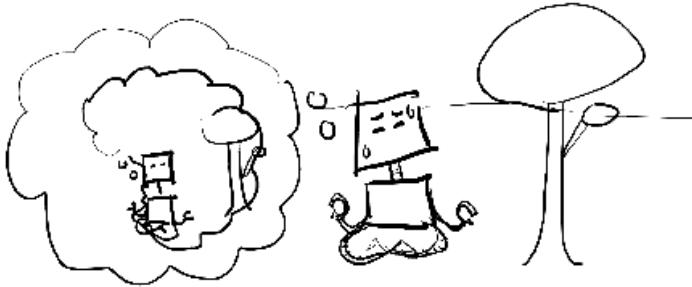
An embedded agent can't do any of those things, at least not in a straightforward way.

One difficulty is that, since it is hard to separate oneself from the world, one would have to have a complete self-model.

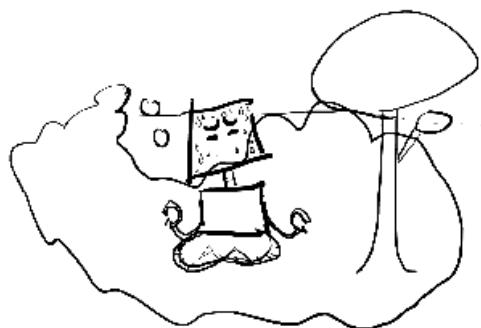




This makes
paradoxes of
self-reference an
obstacle to us.



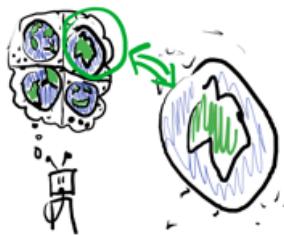
As if representing
the rest of the world
weren't already enough
of a difficulty.



Embedded World-Models have to represent the world in a way more appropriate for embedded agents. Problems in this cluster include:

- world not in hypothesis Space
("realizability"/"grain of truth" problem)
- logical uncertainty
- high-level models
- multi-level models
- ontological crises
- agent must be in world-model (Naturalized Induction)
- anthropic reasoning

In a Bayesian setting, where an agent's uncertainty is quantified by a probability distribution over possible worlds, a common assumption is "realizability": the true underlying environment which is generating the observations is assumed to have at least some probability in the prior.



In game theory, this same property is described by saying a prior has a "grain of truth".

(It should be noted, though, that there are additional barriers to getting this property in a game-theoretic setting; so, in their common usage cases, "grain of truth" is technically demanding while "realizability" is a technical convenience.)

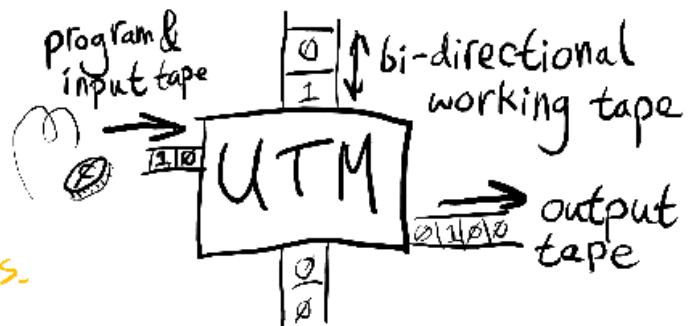
Realizability is not totally necessary in order for Bayesian reasoning to make sense. If you think of a set of hypotheses as "experts", and the current posterior probability as how much you "trust" each expert, then learning according to Bayes' Law ensures a relative bounded loss property.

$$P(h|e) = \frac{P(e|h) P(h)}{P(e)}$$

Specifically, if you use prior π , the amount worse you are in comparison to each expert h is at most $\log(\pi(h))$, since you assign at least probability $\pi(h) \cdot h(e)$ to seeing a sequence of evidence e .

Intuitively, $\pi(h)$ is your initial trust in expert h , and in each case where it is even a little bit more correct than you, you increase your trust accordingly; the way you do this ensures you assign an expert probability 1 and hence copy it precisely before you lose more than $\log(\pi(h))$ compared to it.

The prior AIXI is based on is the Solomonoff prior. It is defined as the output of a universal Turing machine (UTM) whose inputs are coin-flips.



In other words, feed a UTM a random program.

Normally, you'd think of a UTM as only being able to simulate deterministic machines. Here, however, the initial inputs can instruct the UTM to use the rest of the infinite input tape as a source of randomness to simulate a stochastic Turing machine.

Combining this with the previous idea about viewing Bayesian learning as a way of allocating "trust" to "experts" which meets a bounded loss condition, we can see the Solomonoff prior as a kind of ideal machine learning algorithm which can learn to act like any algorithm you might come up with, no matter how clever.

It is assuming all possible algorithms are computable, not that the world is.

For this reason, we shouldn't necessarily think of AIXI as "assuming the world is computable", even though it reasons via a prior over computations.

It's getting bounded loss on its predictive accuracy as compared with any computable predictor.

However, lacking realizability can cause trouble if you are looking for anything more than bounded-loss predictive accuracy.

- posterior can oscillate forever
- probabilities may not be calibrated
- estimates of statistics such as the mean may be arbitrarily bad
- estimates of latent variables may be bad
- identification of causal structure may not work

So, does AIXI perform well without a realizability assumption?

We don't know. Despite getting bounded loss for predictions without realizability, existing optimality results for its actions require a realizability assumption.

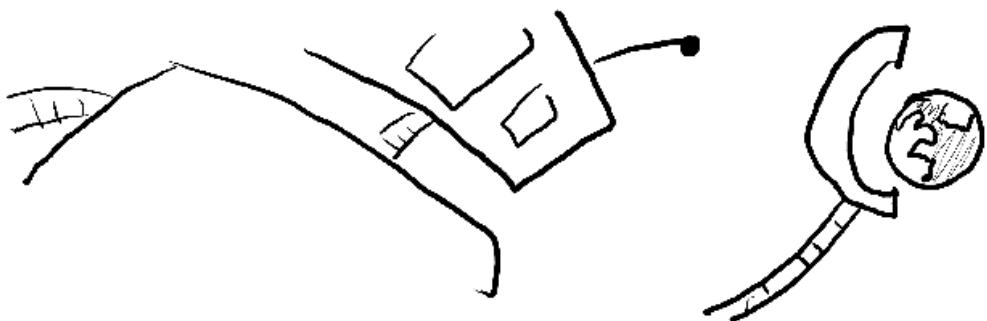
First, if the environment really is sampled from the Solomonoff distribution, AIXI gets the maximum expected reward. But, this is fairly trivial; it is essentially the definition of AIXI.

Second, if we modify AIXI to take somewhat randomized actions (Thompson sampling), there is an asymptotic optimality result for environments which act like any stochastic Turing machine. So, either way, realizability was assumed in order to prove anything.

(Jan Leike, Nonparametric General Reinforcement Learning)

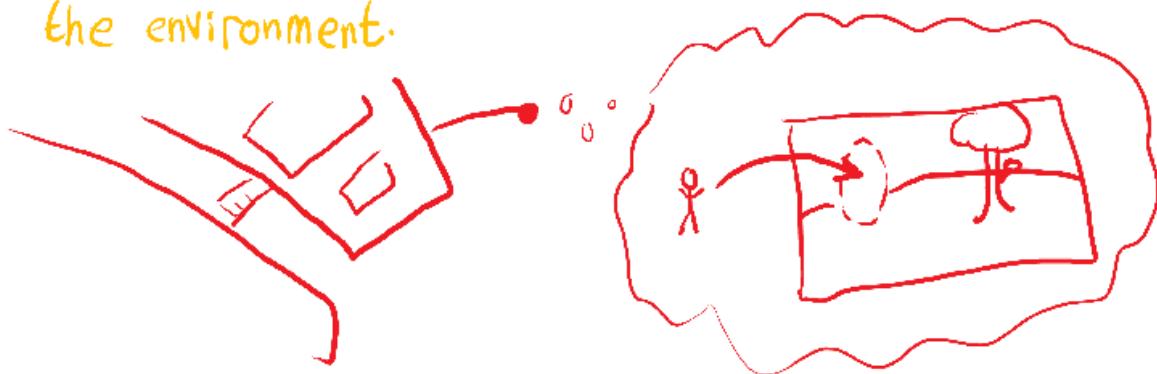
But, the concern I'm pointing at is not "the world might be uncomputable, so we don't know if AIXI will do well" -- that is more of an illustrative case.

The concern is that AIXI is only able to define intelligence/rationality by constructing an agent much much bigger than the environment which it has to learn about and act within.

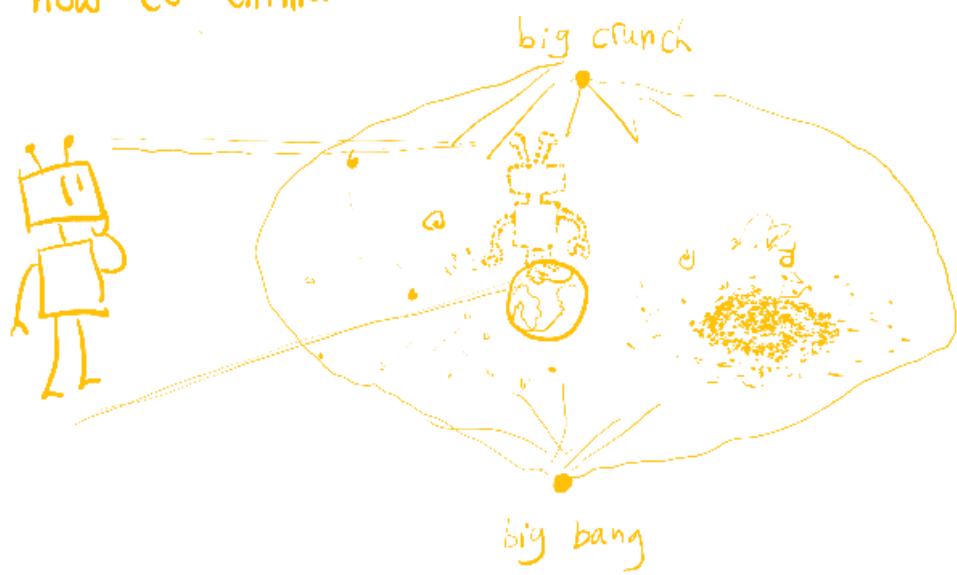


Laurent Orseau provides a way of thinking about this in Space-Time Embedded Intelligence.

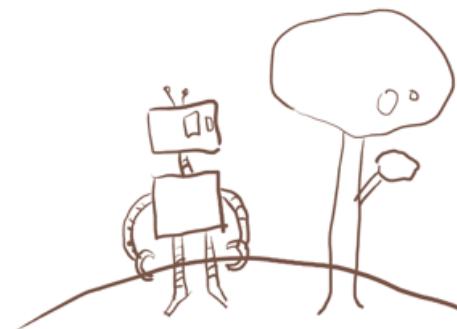
However, his approach defines the intelligence of an agent in terms of a sort of super-intelligent designer who thinks about reality from outside, selecting an agent to place into the environment.



Embedded agents don't have the luxury
of stepping outside of the universe to think
about how to think.



What we would like would be a theory of rational belief for embedded agents which provides similarly strong foundations as Bayesianism provides for Dualistic agents.



$$P(h|e) = \frac{P(e|h) P(h)}{P(e)}$$

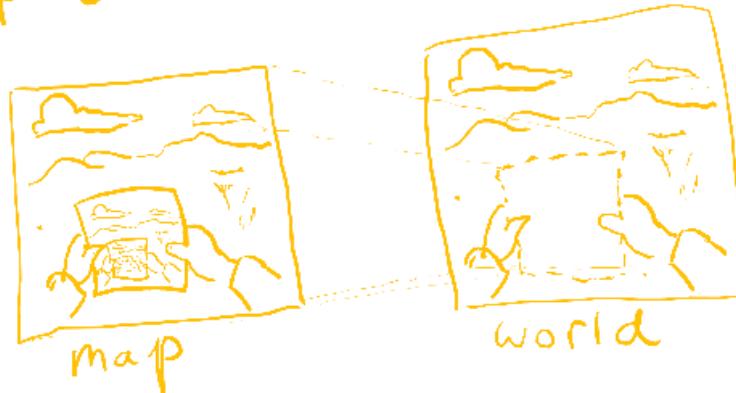


One major obstacle such a theory must deal with is self-reference.

Paradoxes of self-reference such as the Liar paradox make it not just wildly impractical, but in a certain sense impossible for an agent's world-model to accurately reflect the world.

The Liar paradox concerns the status of the self-referential sentence "This sentence is not true.". If it were true, it must be false; and if not true, it must be true.

The difficulty comes in part from trying to draw a map of territory which includes the map itself.



This is fine if the world "holds still" for us; but, because the map is in the world, different maps create different worlds.

We can set up a Liar-paradox-like situation by supposing that we're trying to make an accurate map of the final route of a road which is currently under construction, but we know the construction will proceed so as to disprove whatever map we make.

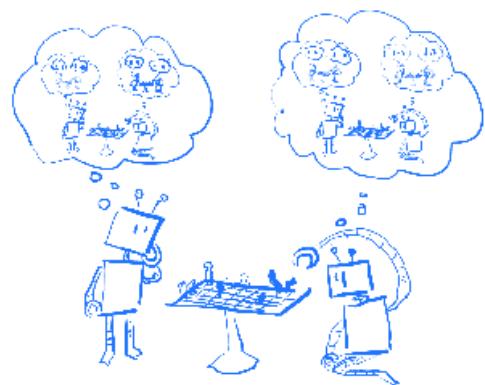


Problems of this kind become relevant for decision-making in the theory of games; a simple game of rock-paper-scissors sets up a Liar paradox, to the extent that the players can predict each other better than chance and are trying to win.

Game theory solves this type of problem with game-theoretic equilibria. But, as I'll explain, realizability issues end up coming back in a different way.

Grain of Truth

I mentioned that the problem of realizability takes on a different character in the context of game theory.



In a machine learning setting, realizability is a potentially unrealistic assumption, but can usually be assumed consistently nonetheless.

Due to the paradoxes which can so easily arise from games, the assumption itself may be inconsistent.

Because there are many agents, it is no longer possible to conveniently make an "agent" a thing which is larger than a world.

So, game-theorists are forced to investigate notions of rational agency which can handle a world which is large.

Unfortunately, this is done by splitting up the world into "agent" parts and "non-agent" parts, and handling the agents in a special way.



This is almost as bad as dualistic models of agency.



In rock-paper-scissors, the Liar paradox is resolved by stipulating that each player play each move with $\frac{1}{3}$ probability. If one player plays this way, then the other loses nothing by doing so.

This way of introducing probabilistic play to resolve would-be paradoxes of game theory is called a Nash equilibrium.

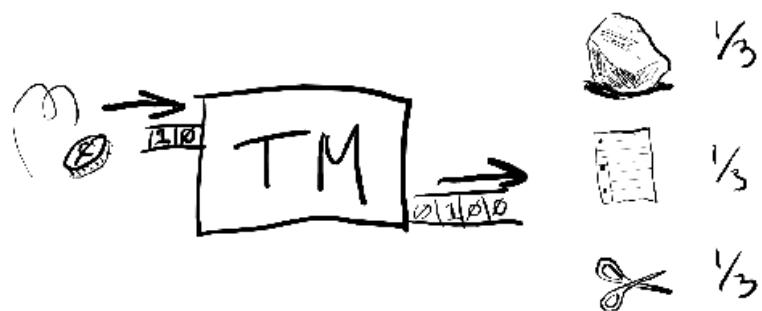
We can use Nash equilibria to prevent the assumption that the agents correctly understand the world they're in from being inconsistent. However, that works just by telling the agents what the world looks like. What if we want to model agents who learn about the world, more like AIXI?

The grain of truth problem is the problem of formulating a reasonably broad prior probability distribution which would allow agents playing games to place some positive probability on each other's true (probabilistic) behavior, without knowing it precisely from the start.

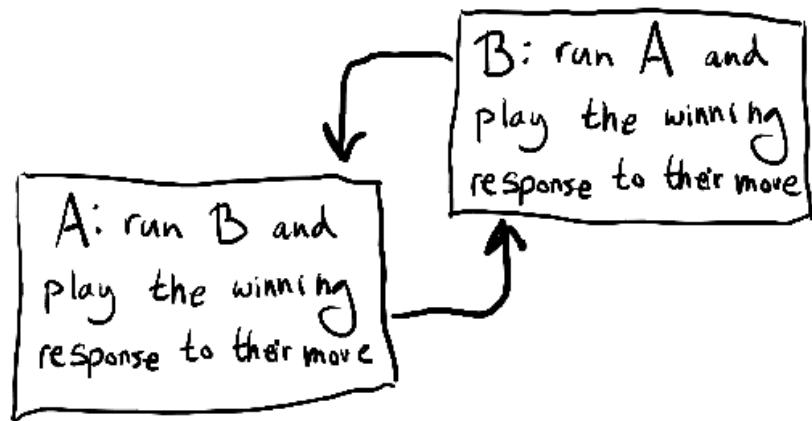


Until recently, known solutions to the problem were quite limited. Reflective Oracles: A Foundation for Classical Game Theory (Benja Fallenstein, Jessica Taylor, & Paul F. Christiano) provides a very general solution (also see A Formal Solution to the Grain of Truth Problem by Jan Leike, Jessica Taylor, & Benja Fallenstein).

You might think stochastic Turing machines can represent Nash equilibria just fine. But, if you're trying to produce them as a result of reasoning about other agents, you'll run into trouble.



If both agents model each other's computation and try to run it to see what each other do, you've just got an infinite loop.



There are some questions Turing machines just can't answer; especially, about the behavior of Turing machines.

The halting problem is the classic example.

Turing studied "oracle machines" to examine what would happen if we could answer such questions.

An oracle is like a book containing some answers to questions which we were unable to answer before.

But ordinarily, we get a hierarchy: type B machines have the answers about type A; type C machines have the answers about types A and B (and so on); but, no machines have answers about their own type.

The diagram illustrates the Oracle Argument through three levels of machines:

- Type A machines (represented by a computer icon) know about "Ah. B-Halting".
- Type B machines (represented by a computer icon) know about "Ah. Halting" and "B-halting??".
- Type C machines (represented by a computer icon) know about "C-halting??", "Ah. B-Halting", and "B-halting".

Ellipses above the machines indicate an infinite hierarchy.

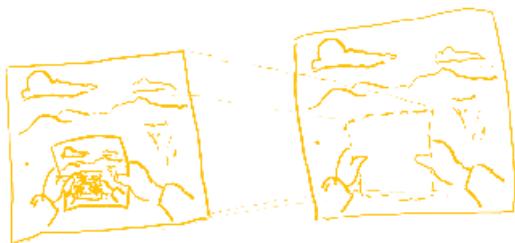


Reflective oracles twist the Turing universe in on itself to avoid this hierarchy, answering questions about the behavior of machines with access to the very same oracle.

The usual paradoxes which this would create are avoided by adding a small amount of randomness, to fuzz things out if e.g. someone tries to do the opposite of what they themselves do.

So R.O. machines are stochastic, but, they're more powerful than regular stochastic Turing machines.



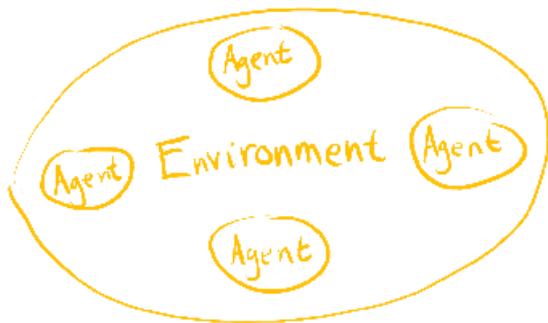


That's how R.O.s address the earlier-mentioned problems of a map that's itself part of the territory:
randomize.



Reflective oracles also solve the problem with game-theoretic notions of rationality I mentioned earlier: it allows agents to be reasoned about in the same manner as other parts of the environment, rather than being a special case to be handled differently.

They're all just computations-with-oracle-access.



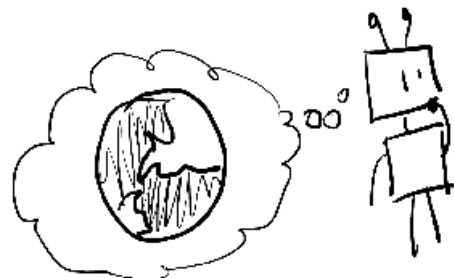
However, models of rational agents based on reflective oracles still have several major limitations. One of these is that agents are required to have infinite processing power, just like AIXI, and so are assumed to know all of the consequences of their own beliefs.

In fact, knowing all the consequences of your beliefs -- a property known as logical omniscience -- turns out to be rather core to classical Bayesian rationality.

Logical Uncertainty

So far, I've been talking in a fairly naive way about the agent having beliefs about hypotheses, and the real world being (or not being) in the hypothesis space.

It isn't really clear what any of that means.



Depending on definitions, it may actually be quite possible for an agent to be smaller than the world and yet "contain" the right world-model — it might know the true physics and initial conditions, but only be capable of inferring their consequences very approximately.



Realistic as this scenario may be, it is not in line with what it usually means for a Bayesian to know something — a Bayesian knows the consequences of all its beliefs.

$$\{P(A)=1\} + A \text{ implies } B \Rightarrow \{P(B)=1\}$$


Uncertainty about the consequences of your beliefs is logical uncertainty. We would like some notion of boundedly rational beliefs about uncertain consequences.



This requires a combined theory of logic — reasoning about implications — and probability — degrees of belief.

Logic and probability theory
are two great triumphs in the
codification of rational thought.

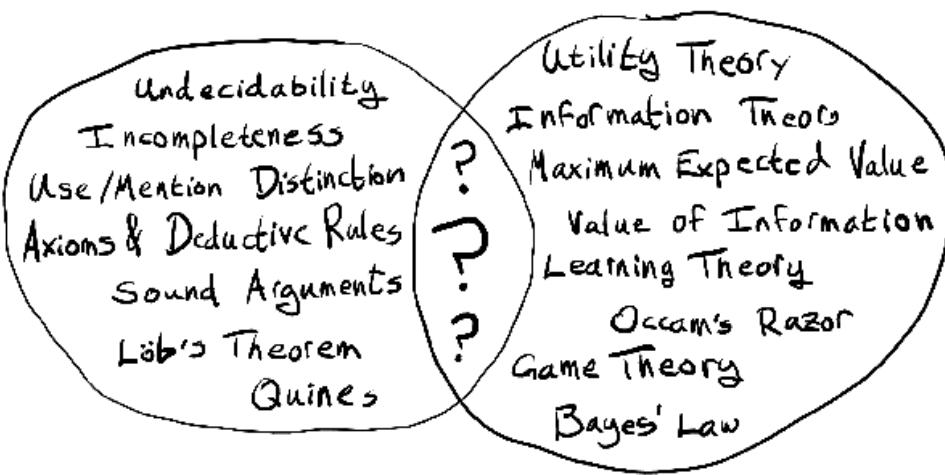
Logic provides
the best tools for
thinking about
self-reference.

Undecidability
Incompleteness
use/Mention Distinction
Axioms & Deductive Rules
Sound Arguments
Löb's Theorem
Quines

Probability provides the
best tools for thinking
about decision-making.

Utility Theory
Information Theory
Maximum Expected Value
Value of Information
Learning Theory
Occam's Razor
Game Theory
Bayes' Law

However, the two don't work together
as well as you may think.



They may seem superficially compatible,
since probability theory is an extension
of Boolean logic.

conjunction $P(A \& B)$

disjunction $P(A \text{ or } B) = P(A) + P(B) - P(A \& B)$

negation $P(\text{not } A) = 1 - P(A)$

However, Gödel's first incompleteness theorem shows, any sufficiently rich logical system is incomplete — ie., not only does it fail to decide every sentence as true or false, it further has no computable extension which manages to do so.

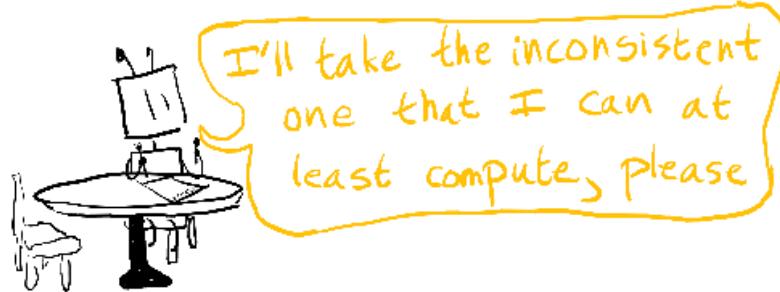
(See the post An Untrollable Mathematician Illustrated for more illustration of how this messes with probability theory.)

This also applies to probability distributions: no computable distribution can assign probabilities in a way that's consistent with a sufficiently rich theory.

This forces us to choose:

- use an uncomputable distribution,
or,
- use one which is inconsistent.

Sounds like an easy choice, right?



I'll take the inconsistent
one that I can at
least compute, please

After all, we're trying to develop a theory of logical non-omniscience here.

We can just continue to update on facts which we prove, bringing us closer and closer to consistency.

Unfortunately, this doesn't work out so well, for reasons which connect back to realizability.



Remember, there are no computable probability distributions consistent with all consequences of rich theories.

So, our non-omniscient prior doesn't even contain a hypothesis which is correct.

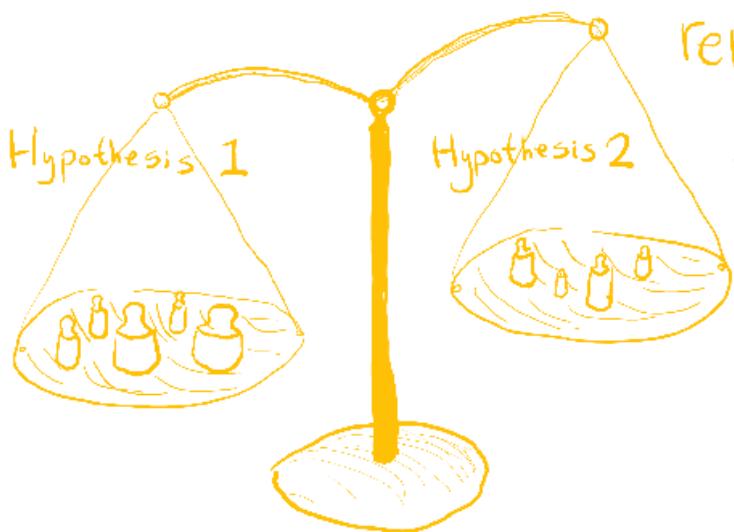
This causes pathological behavior as we condition on more and more true mathematical facts. Beliefs wildly oscillate rather than approaching reasonable estimates.



$$\begin{aligned}1+1 &= 2 \\2+1 &= 3 \\1+2 &= 3 \\2+2 &= 4 \\1+3 &= 4\end{aligned}$$

Taking a Bayesian prior on mathematics, and updating on whatever we prove, does not seem to capture mathematical intuition and heuristic conjecture very well — unless we restrict the domain and craft a sensible prior.

Probability is like a scale, with worlds as weights. An observation eliminates some of the possible worlds, removing weights and shifting the balance of beliefs.



Logic is like a tree, growing from the seed of axioms.



The process of growth is never complete; you never know all the consequences of each belief.

Not knowing the consequences of a belief is like not knowing where to place the weights on the scales of probability.



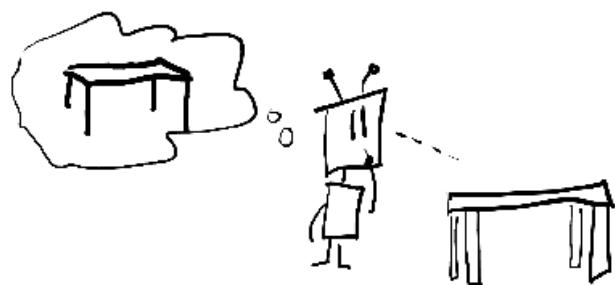
If we put weight in both places until a proof rules one out, the beliefs just oscillate forever rather than doing anything useful.

This grapples directly with the problem of a world which is larger than you.

Any computable beliefs about logic must have left out something, since the tree will grow larger than any container.

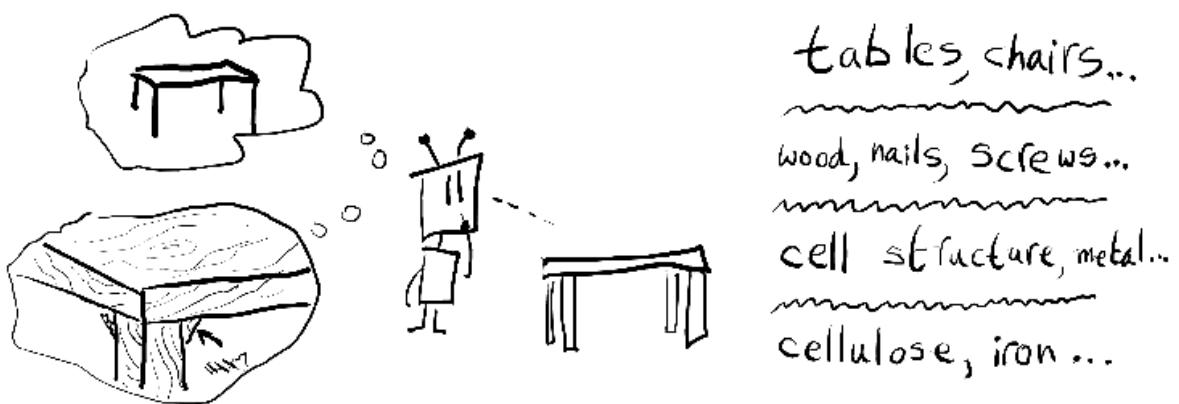


High-Level / Multi-Level World-Models



Because the world is bigger than you, you need to be able to use high-level world models: models which involve things like tables and chairs.

This is related to the classical symbol grounding problem; but, since we want a formal analysis which increases our trust in some system, the kind of model which interests us is somewhat different. This also relates to transparency and informed oversight: world-models should be made out of understandable parts.



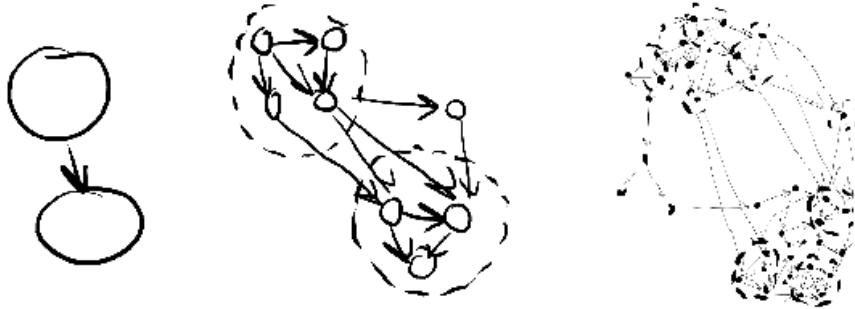
tables, chairs...

wood, nails, screws...

cell structure, metal...

cellulose, iron...

Relatedly, there is the question of how high-level reasoning and low-level reasoning relate to each other and to intermediate levels: multi-level world models. Standard probabilistic reasoning doesn't provide a very good account of this sort of thing.



It's like you've got different Bayes nets which describe the world at different levels of accuracy, and processing power limitations force you to mostly use the less accurate ones, so you have to decide how to jump to the more accurate as needed. Also, the models at different levels don't

line up perfectly, so you have a problem of translating between them. Also, the models may have serious contradictions between them. This might be fine, since high-level models are understood to be approximations anyway; or, it could signal a serious problem in the higher- or lower-level models, requiring their revision.

This is especially interesting in the case of ontological crisis, in which

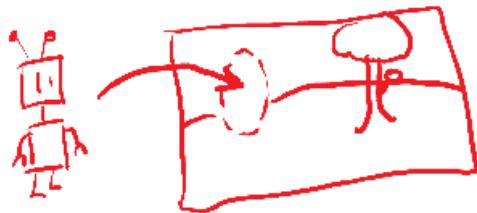
Objects in which we place value turn out not to be a part of "better" models of the world. It seems fair to say that everything humans value exists in high-level models only, which from a reductionistic perspective is "less real" than atoms and quarks.

However, because our values aren't defined on the low level, we are able to keep our values even when our knowledge of

the low level radically shifts. We would also like to be able to say something about what happens to values if the high level radically shifts.

Another critical aspect of embedded world models is that the agent itself must be in the model, since the agent seeks to understand the world, and the world cannot be fully separated from oneself. This opens the door to

difficult problems of self-reference and anthropic decision theory. Naturalized Induction is the problem of learning world-models which include yourself in the environment. This is challenging because (as Caspar Oesterheld has put it) there is a type mismatch between "mental stuff" and "physics stuff".

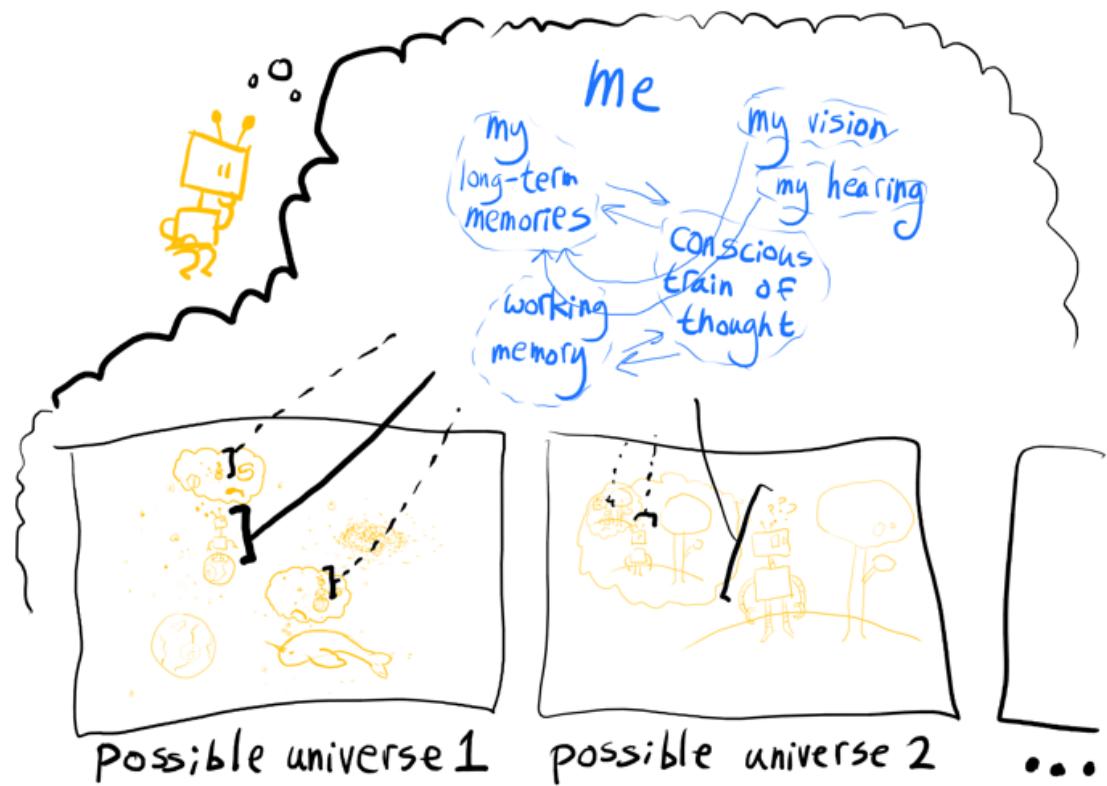


AIXI conceives of the environment as if it were made with a slot which the agent fits into.

We might intuitively reason in this way, but, we can also understand a physical perspective from which this looks like a bad model.

We might imagine instead that the agent separately represents:

- self-knowledge available to introspection
- hypotheses about what the universe is like
- a "bridging hypothesis" connecting the two



There are interesting questions of how this could work. There's also the question of whether this is the right structure at all. It's certainly not how I imagine babies learning.

Thomas Nagel would say this way of approaching the problem involves "views from nowhere"; each hypothesis posits a world as if seen from outside. This is perhaps a strange thing to do.



A particularly interesting aspect of the agent being in its own model of the environment is its need to reason about its future selves. There is a particular sort of trust an agent needs in its future selves to carry out plans; or, because the future is large, to

go beyond plans which can be conceived of at the moment — to pursue one's goals in a reasonable manner.

This is the subject of the next section, robust delegation.

The Inspection Paradox is Everywhere

This is a linkpost for <https://allendowney.blogspot.com/2015/08/the-inspection-paradox-is-everywhere.html?fbclid=IwAR30wr-DdUA6LXEmz4ZfI0IAv9LqUBHUATKY154THEN-SpoYSJgJqvamUuc>

This post describes situations where the average of an attribute observed by a participant is larger than the average over all elements.

Specification gaming examples in AI

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

This is a linkpost for <https://docs.google.com/spreadsheets/u/1/d/e/2PACX-1vRPiprOaC3HsCf5Tuum8bRfzYUiKLRqJmbOoC-32JorNdfyTiRRsR7Ea5eWtvWzuxo8bjOxCG84dAg/pubhtml>

Interesting list of examples where AI programs gamed the specification, solving the problem in rather creative (or dumb) ways not intended by the programmers.

Clickbait might not be destroying our general Intelligence

Epistemic status This post is a "plausible conjecture" alternative to Eliezer's take on clickbait. My arrogant epistemology thinks it's substantially closer to the truth than Eliezer's version.

<https://www.lesswrong.com/posts/YicoiQurNBxSp7a65/is-clickbait-destroying-our-general-intelligence>

In the 1950's there were a similar number (order of magnitude) of humans, and they did a similar amount of socializing as today. This would suggest that a similar quantity of memetic optimization was going on then. Any universally popular meme would have risen to popularity. The difference between then and today was that then people usually only communicated with people geographically local to them. With a small number of newspapers being widely read. Whereas today, people communicate with like minded individuals around the world regularly.

Suppose that different humans have different selection criteria when deciding to share a meme. In 1950, a meme had to appeal to a broad section of society to be spread. Not 100%, more like 25%, the small number of different newspapers, and small numbers of local people to communicate with still let memes specialize along a few socioeconomic lines. Eg Catholic memes vs Protestant memes in Northern Ireland, or Liberal vs Conservative memes in many places. Each newspaper had a side, and many of your neighbors were on your side.

Nowadays, memes can specialize to focus onto tiny subsets of the population. Given that a tiny fraction of the population think in a particular and unusual way, they can gather together online, and share memes optimized exclusively to them. This produces many internet subcultures. Within each subculture, the selection pressure isn't that strong, there aren't that many model railway buffs or code golfers or ... But the memes are optimized to a particular way of thinking, not the combination of several.

In some circumstances, Schelling points create an averaging effect in the 1950 case. Suppose an issue, say cats vs dogs, is sufficiently minor that newspapers are not split into a pro cat paper, and a pro dog paper. Then if a newspaper says anything excessively pro cat, the dog fans will call them out on it, and possibly stop reading the paper. Likewise if the paper is pro dog. So the paper finds a Schelling point, where neither side are upset enough to cause a real fuss. Note that this process is only truth seeking to the extent that cat supporters will let valid pro dog arguments slide and call out invalid ones.

The main effect of filter bubbles is increasing the variance in the meme pool. The biologists can stop arguing with creationists, and get down to sorting out the details of kin selection or whatever. The creationists can stop having to peddle creationism to the unconvincing and can get together to work out the difference between micro-evolution and macro-evolution.

Optimizing for a combination of A and B can produce more of both than choosing which to optimize for at random. If we have memes that are (A=10, B=0) and (A=9, B=9) and (A=0, B=10), then the middle one is likely to spread in a world without filter

bubbles, but the extremes could spread in bubbles optimizing only A and B respectively. If A=sanity, then the average sanity could fall due to the optimization for sanity being focused into one place, and diminishing marginal returns on sanity for optimization.

Other effects on the quality of discourse could include stupid people having an easier time voicing their opinions. Eliezer says that the quality of internet discussion has degraded from 2002 to 2017. (I wasn't old enough to use the internet in 2002, so can't confirm or deny this.) According to these sources, under 10% of the world was online in 2002, facebook and twitter hadn't started yet. In short, getting on the internet required more technical competence, the hardware was more expensive, and there was less to do there. The typical internet user was moderately well educated and smart. The typical newspaper journalist was also moderately well educated. Any reduction in quality would seem to be from uninformed people being finally able to tell the world why the earth is flat.

If you think that the world needs a few highly sane people, not many slightly sane people, then an aggregation into a few groups of sanity is beneficial.

Under the hypothesis that clickbait is destroying intelligence, the existance of less wrong, and places like it, is highly surprising, under a segregation hypothesis, its expected that the most rational people clump together.

<https://www.internetworldstats.com/emarketing.htm>

<https://www.inquisitr.com/830664/the-history-of-social-media-when-did-it-really-begin-you-may-be-surprised-infographic/>

Alignment Newsletter #34

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

Find all Alignment Newsletter resources [here](#). In particular, you can [sign up](#), or look through the [database](#) of all summaries.

Highlights

Scalable agent alignment via reward modeling (*Jan Leike*): This blog post and the [associated paper](#) outline a research direction that DeepMind's AGI safety team is pursuing. The key idea is to learn behavior by learning a reward and a policy simultaneously, from human evaluations of outcomes, which can scale to superhuman performance in tasks where evaluation is easier than demonstration. However, in many cases it is hard for humans to evaluate outcomes: in this case, we can train simpler agents using reward modeling that can assist the human in evaluating outcomes for the harder task, a technique the authors call recursive reward modeling. For example, if you want to train an agent to write a fantasy novel, it would be quite expensive to have a human evaluate outcomes, i.e. rate how good the produced fantasy novels are. We could instead use reward modeling to train agents that can produce plot summaries, assess prose quality and character development, etc. which allows a human to assess the fantasy novels. There are several research challenges, such as what kind of feedback to get, making it sufficiently sample efficient, preventing reward hacking and unacceptable outcomes, and closing the reward-result gap. They outline several promising approaches to solving these problems.

Rohin's opinion: The proposal sounds to me like a specific flavor of narrow value learning, where you learn reward functions to accomplish particular tasks, rather than trying to figure out the "true human utility function". The recursive aspect is similar to [iterated amplification](#) and [debate](#). Iterated amplification and debate can be thought of as operating on a tree of arguments, where each node is the result of considering many child nodes (the considerations that go into the argument). Importantly, the child nodes are themselves arguments that can be decomposed into smaller considerations. Iterated amplification works by learning how to compose and decompose nodes from children, while debate works by having humans evaluate a particular path in the argument tree. Recursive reward modeling instead uses reward modeling to train agents that can help *evaluate outcomes* on the task of interest. This seems less recursive to me, since the subagents are used to evaluate outcomes, which would typically be a different-in-kind task than the task of interest. This also still requires the tasks to be fast -- it is not clear how to use recursive reward modeling to eg. train an agent that can teach math to children, since it takes days or months of real time to even produce outcomes to evaluate. These considerations make me a bit less optimistic about recursive reward modeling, but I look forward to seeing future work that proves me wrong.

The post also talks about how reward modeling allows us to separate what to do (reward) from how to do it (policy). I think it is an open question whether this is desirable. [Past work](#) found that the reward generalized somewhat (whereas policies typically don't generalize at all), but this seems relatively minor. For example, rewards inferred using deep variants of inverse reinforcement learning often don't generalize.

Another possibility is that the particular structure of "policy that optimizes a reward" provides a useful inductive bias that makes things easier to learn. It would probably also be easier to inspect a specification of "what to do" than to inspect learned behavior. However, these advantages are fairly speculative and it remains to be seen whether they pan out. There are also practical advantages: any advances in deep RL can immediately be leveraged, and reward functions can often be learned much more sample efficiently than behavior, reducing requirements on human labor. On the other hand, this design "locks in" that the specification of behavior must be a reward function. I'm not a fan of reward functions because they're so unintuitive for humans to work with -- if we could have agents that work with natural language, I suspect I do not want the natural language to be translated into a reward function that is then optimized.

Technical AI alignment

Iterated amplification sequence

[Prosaic AI alignment](#) (*Paul Christiano*): It is plausible that we can build "prosaic" AGI soon, that is, we are able to build generally intelligent systems that can outcompete humans without qualitatively new ideas about intelligence. It seems likely that this would use some variant of RL to train a neural net architecture (other approaches don't have a clear way to scale beyond human level). We could write the code for such an approach right now (see [An unaligned benchmark](#) from [AN #33](#)), and it's at least plausible that with enough compute and tuning this could lead to AGI. However, this is likely to be bad if implemented as stated due to the standard issues of reward gaming and Goodhart's Law. We do have some approaches to alignment such as IRL and executing natural language instructions, but neither of these are at the point where we can write down code that would plausibly lead to an aligned AI. This suggests that we should focus on figuring out how to align prosaic AI.

There are several reasons to focus on prosaic AI. First, since we know the general shape of the AI system under consideration, it is easier to think about how to align it (while ignoring details like architecture, variance reduction tricks, etc. which don't seem very relevant currently). Second, it's important, both because we may actually build prosaic AGI, and because even if we don't the insights gained will likely transfer. In addition, worlds with short AGI timelines are higher leverage, and in those worlds prosaic AI seems much more likely. The main counterargument is that aligning prosaic AGI is probably infeasible, since we need a deep understanding of intelligence to build aligned AI. However, it seems unreasonable to be confident in this, and even if it is infeasible, it is worth getting strong evidence of this fact in order change priorities around AI development, and coordinate on not building an AGI that is too powerful.

Rohin's opinion: I don't really have much to say here, except that I agree with this post quite strongly.

[Approval-directed agents: overview](#) and [Approval-directed agents: details](#) (*Paul Christiano*): These two posts introduce the idea of approval-directed agents, which are agents that choose actions that they believe their operator Hugh the human would most approve of, if he reflected on it for a long time. This is in contrast to the traditional approach of goal-directed behavior, which are defined by the *outcomes* of the action.

Since the agent Arthur is no longer reasoning about how to achieve outcomes, it can no longer outperform Hugh at any given task. (If you take the move in chess that Hugh most approves of, you probably still lose to Gary Kasparov.) This is still better than Hugh performing every action himself, because Hugh can provide an expensive learning signal which is then distilled into a fast policy that Arthur executes. For example, Hugh could deliberate for a long time whenever he is asked to evaluate an action, or he could evaluate very low-level decisions that Arthur makes billions of times. We can also still achieve superhuman performance by bootstrapping (see the next summary).

The main advantage of approval-directed agents is that we avoid locking in a particular goal, decision theory, prior, etc. Arthur should be able to change any of these, as long as Hugh approves it. In essence, approval-direction allows us to delegate these hard decisions to future overseers, who will be more informed and better able to make these decisions. In addition, any misspecifications seem to cause graceful failures -- you end up with a system that is not very good at doing what Hugh wants, rather than one that works at cross purposes to him.

We might worry that *internally* Arthur still uses goal-directed behavior in order to choose actions, and this internal goal-directed part of Arthur might become unaligned. However, we could even have internal decision-making about cognition be approval-based. Of course, eventually we reach a point where decisions are simply made -- Arthur doesn't "choose" to execute the next line of code. These sorts of things can be thought of as heuristics that have led to choosing good actions in the past, that could be changed if necessary (eg. by rewriting the code).

How might we write code that defines approval? If our agents can understand natural language, we could try defining "approval" in natural language. If they are able to reason about formally specified models, then we could try to define a process of deliberation with a simulated human. Even in the case where Arthur learns from examples, if we train Arthur to predict approval from observations and take the action with the highest approval, it seems possible that Arthur would not manipulate approval judgments (unlike AIXI).

There are also important details on how Hugh should rate -- in particular, we have to be careful to distinguish between Hugh's beliefs/information and Arthur's. For example, if Arthur thinks there's a 1% chance of a bridge collapsing if we drive over it, then Arthur shouldn't drive over it. However, if Hugh always assigns approval 1 to the optimal action and approval 0 to all other actions, and Arthur believes that Hugh knows whether the bridge will collapse, then the maximum expected approval action is to drive over the bridge.

The main issues with approval-directed agents is that it's not clear how to define them (especially from examples), whether they can be as useful as goal-directed agents, and whether approval-directed agents will have internal goal-seeking behavior that brings with it all of the problems that approval was meant to solve. It may also be a problem if some other Hugh-level intelligence gets control of the data that defines approval.

Rohin's opinion: Goal-directed behavior requires an extremely intelligent overseer in order to ensure that the agent is pointed at the correct goal (as opposed to one the overseer thinks is correct but is actually slightly wrong). I think of approval-directed agents as providing the intuition that we may only require an overseer that is slightly smarter than the agent in order to be aligned. This is because the overseer can simply

"tell" the agent what actions to take, and if the agent makes a mistake, or tries to optimize a heuristic too hard, the overseer can notice and correct it interactively. (This is assuming that we solve the [informed oversight problem](#) so that the agent doesn't have information that is hidden from the overseer, so "intelligence" is the main thing that matters.) Only needing a slightly smarter overseer opens up a new space of solutions where we start with a human overseer and subhuman AI system, and scale both the overseer and the AI at the same time while preserving alignment at each step.

[Approval-directed bootstrapping](#) (*Paul Christiano*): To get a very smart overseer, we can use the idea of bootstrapping. Given a weak agent, we can define a stronger agent that happens from letting the weak agent think for a long time. This strong agent can be used to oversee a slightly weaker agent that is still stronger than the original weak agent. Iterating this process allows us to reach very intelligent agents. In approval-directed agents, we can simply have Arthur ask Hugh to evaluate approval for actions, and *in the process of evaluation* Hugh can consult Arthur. Here, the weak agent Hugh is being amplified into a stronger agent by giving him the ability to consult Arthur -- and this becomes stronger over time as Arthur becomes more capable.

Rohin's opinion: This complements the idea of approval from the previous posts nicely: while approval tells us how to build an aligned agent from a slightly smarter overseer, bootstrapping tells us how to improve the capabilities of the overseer and the agent.

[Humans Consulting HCH](#) (*Paul Christiano*): Suppose we unroll the recursion in the previous bootstrapping post: in that case, we see that Hugh's evaluation of an answer can depend on a question that he asked Arthur whose answer depends on how Hugh evaluated an answer that depended on a question that he asked Arthur etc. Inspired by this structure, we can define HCH (humans consulting HCH) to be a process that answers question Q by perfectly imitating how Hugh would answer question Q, *if Hugh had access to the question-answering process*. This means Hugh is able to consult a copy of Hugh, who is able to consult a copy of Hugh, who is able to consult a copy of Hugh, ad infinitum. This is one proposal for how to formally define a human's enlightened judgment.

You could also combine this with particular ML algorithms in an attempt to define versions of those algorithms aligned with Hugh's enlightened judgment. For example, for RL algorithm A, we could define max-HCH_A to be A's chosen action when maximizing Hugh's approval after consulting max-HCH_A.

Rohin's opinion: This has the same nice recursive structure of bootstrapping, but without the presence of the agent. This probably makes it more amenable to formal analysis, but I think that the interactive nature of bootstrapping (and iterated amplification more generally) is quite important for ensuring good outcomes: it seems way easier to control an AI system if you can provide input constantly with feedback.

Fixed point sequence

[Fixed Point Discussion](#) (*Scott Garrabrant*): This post discusses the various fixed point theorems from a mathematical perspective, without commenting on their importance for AI alignment.

Technical agendas and prioritization

[Integrative Biological Simulation, Neuropsychology, and AI Safety](#) (*Gopal P. Sarma et al*): See [Import AI](#) and [this comment](#).

Learning human intent

[Scalable agent alignment via reward modeling](#) (*Jan Leike*): Summarized in the highlights!

Adversarial examples

[A Geometric Perspective on the Transferability of Adversarial Directions](#) (*Zachary Charles et al*)

AI strategy and policy

[MIRI 2018 Update: Our New Research Directions](#) (*Nate Soares*): This post gives a high-level overview of the new research directions that MIRI is pursuing with the goal of deconfusion, a discussion of why deconfusion is so important to them, an explanation of why MIRI is now planning to leave research unpublished by default, and a case for software engineers to join their team.

Rohin's opinion: There aren't enough details on the technical research for me to say anything useful about it. I'm broadly in support of deconfusion but am either less optimistic on the tractability of deconfusion, or more optimistic on the possibility of success with our current notions (probably both). Keeping research unpublished-by-default seems reasonable to me given the MIRI viewpoint for the reasons they talk about, though I haven't thought about it much. See also [Import AI](#).

Other progress in AI

Reinforcement learning

[Woulda, Coulda, Shoulda: Counterfactually-Guided Policy Search](#) (*Lars Buesing et al*) (summarized by Richard): This paper aims to alleviate the data inefficiency of RL by using a model to synthesise data. However, even when environment dynamics can be modeled accurately, it can be difficult to generate data which matches the true distribution. To solve this problem, the authors use a Structured Causal Model trained to predict the outcomes which would have occurred if different actions had been taken from previous states. Data is then synthesised by rolling out from previously-seen states. The authors test performance in a partially-observable version of SOKOBAN, in which their system outperforms other methods of generating data.

Richard's opinion: This is an interesting approach which I can imagine becoming useful. It would be nice to see more experimental work in more stochastic environments, though.

[Natural Environment Benchmarks for Reinforcement Learning](#) (*Amy Zhang et al*) (summarized by Richard): This paper notes that RL performance tends to be measured in simple artificial environments - unlike other areas of ML in which using real-world data such as images or text is common. The authors propose three new benchmarks to address this disparity. In the first two, an agent is assigned to a random location in an image, and can only observe parts of the image near it. At every time step, it is able to move in one of the cardinal directions, unmasking new sections of the image, until it can classify the image correctly (task 1) or locate a given object (task 2). The third type of benchmark is adding natural video as background to existing Mujoco or Atari tasks. In testing this third category of benchmark, they find that PPO and A2C fall into a local optimum where they ignore the observed state when deciding the next action.

Richard's opinion: While I agree with some of the concerns laid out in this paper, I'm not sure that these benchmarks are the best way to address them. The third task in particular is mainly testing for ability to ignore the "natural data" used, which doesn't seem very useful. I think a better alternative would be to replace Atari with tasks in procedurally-generated environments with realistic physics engines. However, this paper's benchmarks do benefit from being much easier to produce and less computationally demanding.

Deep learning

[Do Better ImageNet Models Transfer Better?](#) (*Simon Kornblith et al*) (summarized by Dan H)

Dan H's opinion: This paper shows a strong correlation between a model's ImageNet accuracy and its accuracy on transfer learning tasks. In turn, better ImageNet models learn stronger features. This is evidence against the assertion that researchers are simply overfitting ImageNet. [Other evidence](#) is that the architectures themselves work better on different vision tasks. Further evidence against overfitting ImageNet is that many architectures which are designed for CIFAR-10, when trained on ImageNet, [can be highly competitive on ImageNet](#).

[Gather-Excite: Exploiting Feature Context in Convolutional Neural Networks](#) (*Jie Hu, Li Shen, Samuel Albanie et al*) (summarized by Dan H)

Read more: This method uses spatial summarization for increasing convnet accuracy and was discovered around the same time as [this similar work](#). Papers with independent rediscoveries tend to be worth taking more seriously.

[Improving Generalization for Abstract Reasoning Tasks Using Disentangled Feature Representations](#) (*Xander Steenbrugge et al*)

What are Universal Inductors, Again?

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

[Attention Conservation Notice: This just recaps an old result and patches a hole in it, it doesn't contain substantially new ideas.]

Universal Inductors can be thought of as logical inductors over bitstrings, which are notable because they can act as a logical inductor over any theory you want (PA, ZFC, 2nd order arithmetic), by fixing some efficiently computable function from bits to sentences in the language of your theory, conditioning the universal inductor (which is a probability distribution over infinite bitstrings) on the bits which correspond to proven theorems, and reading out the probability of other bits/sentences from the resulting conditional distribution. This trick works by Theorem 4.7.2 in the logical induction paper, "Closure Under Conditioning", which shows that conditioning the probabilities of a logical inductor on a sequence of non-inconsistent sentences, yields a logical inductor over a deductive process where all the conditionals are true.

Also, since they are a probability distribution over infinite bitstrings, it makes sense to think about them as having a probability measure over worlds (assignments of statements to true or false, for all statements) at all finite stages, instead of just in the limit.

However, in Scott's [old post about universal inductors](#), their construction was never described.

Note that a Universal Inductor corresponds to a Logical Inductor, but the associated Logical Inductor will not have finite support, and so will be different from the one constructed in the paper. Never the less, Universal Inductors can be shown to exist using a similar construction.

This post will give that construction. To begin with, a Universal Inductor must fulfill the following two properties:

First, it must be a distribution over infinite bitstrings, such that $P_n(\sigma)$ is computable for all n and all σ which are finite bitstrings. This gives the probability that an infinite bitstring starts with σ as a prefix.

Second, the theory is one where the n th atomic statement is of the form "the n th bit in this infinite bitstring is a 1", and P_n induces a function from these statements to probabilities, and the sequence of probability assignments must form a logical inductor over the empty deductive process. (ie, the inductor never sees any evidence of the form "this bit is a 1", it just runs forever without getting any feedback.)

To begin with, the supertrader construction from Section 5 of the [logical induction paper](#) is the exact same. Traders look at the prices of boolean combinations of atomic statements, and use them in continuous functions to buy or sell shares in boolean combinations of atomic statements.

The interesting part is the construction of the space that we'll be finding a fixed-point in. At timestep n , there is a most-distant bit that has ever appeared in a boolean combination of bits that a trader has bought/sold shares in, or looked at the price of. This is bit $f(n)$. If we assign prices to all bitstrings of length $f(n)$, then the probability of any finite bitstring that is longer can be given by just using the uniform distribution on bits after that point. This corresponds to assigning 50/50 probability to all statements which are too large to have thought about them yet. Therefore, our space of interest is the $2^{f(n)} - 1$ dimensional simplex of probability distributions over all bitstrings of length $f(n)$.

Given a point in this space, it is possible to read out the probability of any particular boolean combination of bits by just summing up the probability on all $f(n)$ -length bitstrings that fit that constraint, so we can determine the prices that the supertrader sees.

We can also assign all bitstrings but one to a basis vector for the space, by letting the basis vector for a bitstring be the vector pointing to the corresponding vertex of the simplex from its center. A purchase of one share in a boolean combination corresponds to the vector comprised of the sum of basis vectors that correspond to the bitstrings fulfilling the conditions of the boolean. Note that a purchase of the one bitstring we left out can be re-expressed as a purchase of 1 dollar and selling a share of all other bitstrings.

As a simple case of this re-expression, when $f(n) = 2$, $P(00) = 1 - P(01) - P(10) - P(11)$ (because it has to be a probability distribution), so the price of the purchase is the same. And as for the purchase, both purchases have

the feature that they are worth 1 dollar if the world is 00 and 0 dollars if the world is something different.

Therefore, given a point \mathbf{x} in the simplex, the net purchase and selling of a bunch of different boolean combinations of bits can be broken down into a whole bunch of vectors, which sum up to give a net trade vector \mathbf{t} , and finding the point in the simplex that's closest to $\mathbf{x} + \mathbf{t}$ is the output of the continuous function that we can then find the fixed-point of.

Now, given a fixed-point in the interior of the simplex, $\mathbf{t} = 0$, so there is no net trade. Given a fixed-point on the boundary of the simplex, \mathbf{t} can be interpreted as being composed entirely of selling shares with a price of 0, which obviously has 0 or negative value, although this is quite nontrivial to show. Once this is proved, it then immediately follows that this process described above is a Universal Inductor.

From here on, we will simply take care of this last theorem to be shown.

Theorem 1:

For all price vectors \mathbf{x}' on the boundary of the simplex that are a fixed-point, the vector \mathbf{v} representing the net trade can be interpreted as being composed entirely of 0's for all components in which \mathbf{x}' has a nonzero entry, and nonpositive for all components in which \mathbf{x}' has a 0 entry.

First, some notation will be laid out.

$k := 2^{f(n)} - 1$, and it is the dimension of the space the simplex is in. \mathbf{x} denotes the vector of length k that is the fixed-point, and \mathbf{x}' denotes the unique "canonical interpretation" of this vector as a vector of length $k + 1$ with nonnegative coordinates that sum to 1, which gives the probabilities of the various bitstrings. Similarly, \mathbf{t} denotes the vector of length k which corresponds to the net trade. Given a constant c , \mathbf{e} will denote a vector with all coordinates being c , with the length clear from context. Given a vector \mathbf{z} and coordinate a , with z_a being the value of the a 'th coordinate of \mathbf{z} , \mathbf{z}_a is the vector that is 0 in all coordinates except for a , and z_a at the a 'th coordinate.

To begin, note that, given \mathbf{t} , there are multiple possible vectors \mathbf{t}' of length $k + 1$ which give the net purchase of shares. As an example of this, when $k = 1$, $[0, 1]$ and $[-1, 0]$ both correspond to the same trade, because buying a share of $\sigma = 1$ is equivalent to selling a share of $\sigma = 0$. We will show that \mathbf{t}' has an interpretation as a vector \mathbf{t}' which is 0 everywhere \mathbf{x}' is positive, and nonpositive everywhere \mathbf{x}' is 0, because this corresponds to the net trade being interpretable as buying no shares, and only selling shares of bitstrings with a price of 0, which cannot lead to any gain no matter which bitstring is the true one.

Now, we will fix our basis vectors for the space. Because \mathbf{x} is on the boundary of the simplex, \mathbf{x}' must have at least one 0 entry. Fix that vertex that corresponds to that bitstring as the origin of the coordinate system, and let the basis be comprised of the set of unit vectors that point from the center of the simplex to the vertices that correspond to all the other bitstrings. This spans the space that the simplex is embedded in. Annoyingly, this is not an orthonormal basis, so computing the dot product will be more complicated. The dot product between any two different basis vectors is $\frac{1}{k+1}$ (hat tip to Vanessa from MIRIxDiscord)

We can start by asking the question "In this basis, what condition on a vector \mathbf{z} corresponds to being inside the simplex?" The condition is that there exist a $c \in [0, 1]$ s.t. $[0, \mathbf{z}] + \mathbf{e}$ has all entries in $[0, 1]$ and the entries sum up

to 1. Note that for \mathbf{x} specifically, by the way we have defined the basis, $c = 0$.

Also, the coordinates for \mathbf{x} can be broken down into coordinates where \mathbf{x} has 0 entries, and coordinates where \mathbf{x} has positive entries. Therefore, given an arbitrary vector \mathbf{z} , it can be expressed as $[\mathbf{z}_0, \mathbf{z}_\neq]$, where \mathbf{z}_0 is the vector of coordinates where \mathbf{x}_0 is 0, and \mathbf{z}_\neq is the vector of coordinates where \mathbf{x}_0 is positive.

The proof strategy that will be used is dividing into cases, and showing that for all cases besides the ones that prove the theorem, there is an "allowable perturbation vector" \mathbf{e} s.t. $\mathbf{x} + \mathbf{e}$ lies in the simplex, and $\mathbf{x} + \mathbf{t}^*$ is closer to $\mathbf{x} + \mathbf{e}$ than it is to \mathbf{x} . This implies that \mathbf{x} is not a fixed-point, because it isn't the point closest to $\mathbf{x} + \mathbf{t}^*$, but \mathbf{x} is a fixed-point, so we have a contradiction. Note that $\mathbf{x} + \mathbf{t}^*$ being closer to $\mathbf{x} + \mathbf{e}$ than to \mathbf{x} is equivalent to $\|\mathbf{t}^*\| > \|\mathbf{t}^* - \mathbf{e}\|$, which can be rewritten as $\langle \mathbf{t}^*, \mathbf{t}^* \rangle > \langle \mathbf{t}^*, \mathbf{t}^* \rangle - 2\langle \mathbf{t}^*, \mathbf{e} \rangle + \langle \mathbf{e}, \mathbf{e} \rangle$ which can be rewritten as $2\langle \mathbf{t}^*, \mathbf{e} \rangle > \langle \mathbf{e}, \mathbf{e} \rangle$. This last statement will be the proof target in the following cases.

Case 1: There is some pair of coordinates a, b , s.t $x_b > 0$, and $t_a > t_b$.

Let \mathbf{e} be s.t. $e_a = \epsilon$, $e_b = -\epsilon$, and all other coordinates are 0. Note that since $x_a < 1$ (because $x_b > 0$ and \mathbf{x} adds up to 1), and \mathbf{e} adds up to 0, then for sufficiently small ϵ , $[0, \mathbf{x} + \mathbf{e}]$ has all entries lie in $[0, 1]$ and sums up to 1. Then c can be taken as 0, to yield that $\mathbf{x} + \mathbf{e}$ lies in the simplex.

Define \mathbf{t}^* as $\mathbf{t}^* = \mathbf{t}_0 + \mathbf{t}_\neq + \mathbf{t}^*$. Note that \mathbf{t}^* is 0 at coordinates a and b .

$$2\langle \mathbf{t}^*, \mathbf{e} \rangle = 2(\langle \mathbf{t}_0, \mathbf{e}_a \rangle + \langle \mathbf{t}_0, \mathbf{e}_b \rangle + \langle \mathbf{t}_\neq, \mathbf{e}_a \rangle + \langle \mathbf{t}_\neq, \mathbf{e}_b \rangle + \langle \mathbf{t}^*, \mathbf{e}_a \rangle + \langle \mathbf{t}^*, \mathbf{e}_b \rangle)$$

Rewriting $\langle \mathbf{t}^*, \mathbf{e}_a \rangle + \langle \mathbf{t}^*, \mathbf{e}_b \rangle$ as $\sum_{i \neq a, b} (\langle \mathbf{t}_i^*, \mathbf{e}_a \rangle + \langle \mathbf{t}_i^*, \mathbf{e}_b \rangle)$, and using the fact that the dot product between any two different basis vectors of our space is 0, $\langle \mathbf{t}_i^*, \mathbf{e}_a \rangle + \langle \mathbf{t}_i^*, \mathbf{e}_b \rangle = \epsilon t_i^* (\epsilon t_i - \epsilon t_i) = 0$, so $\langle \mathbf{t}^*, \mathbf{e}_a \rangle + \langle \mathbf{t}^*, \mathbf{e}_b \rangle = 0$.

$$\text{Now, } 2(\langle \mathbf{t}_0, \mathbf{e}_a \rangle + \langle \mathbf{t}_0, \mathbf{e}_b \rangle + \langle \mathbf{t}_\neq, \mathbf{e}_a \rangle + \langle \mathbf{t}_\neq, \mathbf{e}_b \rangle) = 2(\epsilon t_a - \frac{\epsilon}{1-\epsilon} \epsilon t_a + \frac{\epsilon}{1-\epsilon} \epsilon t_b - \epsilon t_b)$$

$$= 2\epsilon(1 + \frac{\epsilon}{1-\epsilon})(t_a - t_b) \text{ and because } t_a > t_b, 2\langle \mathbf{t}^*, \mathbf{e} \rangle = ce \text{ for some positive } c.$$

A similar analysis applies to show $\langle \mathbf{e}, \mathbf{e} \rangle = 4\epsilon^2(1 + \frac{\epsilon}{1-\epsilon}) = de^2$ for some positive d , and for sufficiently small ϵ , we get our desired result that $2\langle \mathbf{t}^*, \mathbf{e} \rangle > \langle \mathbf{e}, \mathbf{e} \rangle$.

Now, the elimination of this case shows that \mathbf{t}_\neq must have all entries equal some constant c , while \mathbf{t}_0 must have all entries $\leq c$ if it is nonempty. This produces three more exhaustive cases.

Case 2: $\mathbf{t}_\neq = \mathbf{e}$, c is positive.

In this case, note that a purchase of all shares but one can be reinterpreted as buying 0 of all shares but one, and selling the remaining share. Therefore, this trade is equivalent to selling c shares in the probability-zero bitstring that we took as the origin, buying 0 of all shares with positive probability, and selling a non-negative amount of all shares with 0 probability (because \mathbf{t}_0 has all entries $\leq c$). Theorem 1 follows.

Case 3: $t_a^* = \epsilon$, c is 0.

This case immediately proves Theorem 1, because $t_a^* = 0$, and t_a^* is 0 or negative on all entries.

Case 4: $t_a^* = \epsilon$, c is negative.

If we can show that this case is impossible, we will be done. Let a be some coordinate where x^* is positive.

Let ϵ be s.t. $\epsilon_a = -2\epsilon$, and all other coordinates are $-\epsilon$. Let $c = \epsilon$. $[0, x^* + \epsilon] + \epsilon$ has all terms lie in $[0, 1]$ for sufficiently small ϵ , because the $-\epsilon$ is canceled out when ϵ is added, and the -2ϵ is taken out of a positive term. Because ϵ has $k+1$ entries, ϵ sums up to $(k+1)\epsilon$. x^* sums up to 1, and ϵ sums up to $-(k+1)\epsilon$, so the resulting vector sums up to 1, and $x^* + \epsilon$ lies in the simplex.

$$2(t^*, \epsilon) = 2((t_a^*, \epsilon_a) + (t_a^*, \epsilon*) + (t^*, \epsilon_a) + (t^*, \epsilon*))$$

Now we will break down the dot products.

$$\begin{aligned} (t^*, \epsilon*) &= \sum_{i,j \neq a} (t_i^*, \epsilon_j^*) = \sum_{i \neq a} ((t_i^*, \epsilon_i^*) + \sum_{j \neq i, a} (t_i^*, \epsilon_j^*)) \\ &= \sum_{i \neq a} (-\epsilon v_i + \sum_{j \neq i, a} (-\epsilon v_i)) = -\epsilon \sum_{i \neq a} (v_i - \frac{1}{k+1} v_i) = -\epsilon (1 - \frac{1}{k+1}) \sum_{i \neq a} v_i \\ &= \frac{k-1}{k+1} \sum_{i \neq a} v_i \end{aligned}$$

And for the next pair,

$$\begin{aligned} (t_a^*, \epsilon*) + (t^*, \epsilon_a) &= \sum_{i \neq a} (t_a^*, \epsilon_i^*) + \sum_{i \neq a} (t_i^*, \epsilon_a) = \sum_{i \neq a} (-\epsilon \frac{1}{k+1} t_a) + \sum_{i \neq a} (-2\epsilon \frac{1}{k+1} t_i) \\ &= \frac{1}{k+1} (k-1)t_a + \frac{1}{k+1} \sum_{i \neq a} t_i \end{aligned}$$

Note that the second term combines with the term from the first dot product we looked at to yield the following equation

$$(t_a^*, \epsilon*) + (t^*, \epsilon_a) + (t^*, \epsilon_a) = \frac{1}{k+1} (k-1)t_a - \frac{1}{k+1} \sum_{i \neq a} t_i$$

Finally, because $(t_a^*, \epsilon_a) = -2\epsilon v_a$, by grouping terms and factoring out $\frac{1}{k+1}$, we get

$$2(t^*, \epsilon) = \frac{1}{k+1} (-2(k+1)t_a + (k-1)t_a - \sum_{i \neq a} t_i) = \frac{1}{k+1} (-3t_a - \sum_{i \neq a} t_i)$$

and, because all coordinates in t^* are negative, this equals ce for some positive constant c . By the same proof path,

$$(\epsilon, \epsilon) = \frac{1}{k+1} (-(k+3)(-2\epsilon) - (k-1)(-\epsilon)) = \frac{1}{k+1} (3k+5) = d\epsilon^2$$

For some positive constant d . Therefore, for sufficiently small ϵ , $2(t^*, \epsilon) > (\epsilon, \epsilon)$ and Case 4 is impossible, and the theorem follows because only Case 2 and 3 are left.

Update the best textbooks on every subject list

I occasionally refer back to lukeprog's [Best Textbooks on Every Subject](#) post. I thought it might be a good idea to direct people back to it in the hopes of updating the list, for the following reasons:

- Old lesswrong is fully integrated now, so we can do it from this site.
- It hasn't been significantly updated in 8 years; since then it seems like the community has both diversified and increased in size, so hopefully we can both broaden and deepen the list.
- In the interim there have been several rounds of people who have done various levels of [MIRI's research guide](#), and it seems like there is richer engagement with the fields surrounding rationality.
- A lot of textbooks get published in 8 years; new ones may be improvements over the old, or we may have gained textbooks for fields which lacked them previously.
- I would be particularly interested in anything which accounts for the replication crisis, especially with respect to important fields like behavioral economics.

At ChristianKI's suggestion:

Here are **the rules:**

1. Post the title of your favorite textbook on a given subject.
2. You must have read at least two other textbooks on that same subject.
3. You must briefly name the other books you've read on the subject and explain why you think your chosen textbook is superior to them.

Burnout: What it is and how to Treat it.

This is a linkpost for

<https://forum.effectivealtruism.org/posts/NDszJWMsdLCB4MNoy/burnout-what-is-it-and-how-to-treat-it>

I reviewed the scientific literature on burnout to create a better definition and treatment plan.

TL;dr

Social support == Good.

Sleep == Good.

Ambiguity == Bad.

Vacations == Meh.

I expect a lot of value in the comments and want to keep them all in one place, so comments will be disabled here.

How democracy ends: a review and reevaluation

Last month I attended a talk by David Runciman, the author of a recent book called *How Democracy Ends*. I was prepared for outrage-stirring and pearl-clutching, but was pleasantly surprised by the quality of his arguments, which I've summarised below, along with my own thoughts on these issues. Note, however, that I haven't read the book itself, and so can't be confident that I've portrayed his ideas faithfully.

Which lessons from history?

Many people have compared recent populist movements with the stirrings of fascism a century ago. And indeed it's true that a lot of similar rhetoric being thrown around. But Runciman argues that this is one of the least interesting comparisons to be made between these two times. Some things that would be much more surprising to a denizen of the early 20th century:

- Significant advances in technology
- Massive transformations in societal demographics
- Very few changes in our institutions

The last of those is particularly surprising in light of the first two. Parliamentary democracies in the Anglophone world have been governed by the same institutions - and in some cases, even the same parties - for centuries. Continental European democracies were more disrupted by World War 2, but have been very stable since then, despite the world changing in many ways overall. That's true even for institutions that are probably harmful - consider the persistence of the electoral college in the US, the House of Lords in the UK, the monarchy in Australia (despite their ardent republicanism movement), first-past-the-post voting systems in many countries, and so on. (In fact, Runciman speculates that Americans voted for Trump partly because of how much confidence they had in the durability of their institutions - a confidence which so far seems to have been well-founded.)

So history gives us pretty good evidence for the robustness of democratic institutions to the normal flow of time - but not to exceptional circumstances. In fact, an inability to make necessary changes may well render them more fragile in the face of sharp opposition. If and when pressure mounts, are they going to snap like the democratic institutions of 1930s Europe did? Runciman argues that they won't, because of the nature of the demographic changes the West has seen. There are three particularly important axes of variation:

- Wealth: the average person is many times wealthier than they were a century ago, and the middle class is much larger.
- Education: we've gone from only a few percent of people getting tertiary education (and many of the remainder not finishing high school) to nearly 50% of young people being in university in many western countries.
- Age: in the last century, the median age has risen by over ten years in most western countries.

These three factors are some of the most powerful predictors of behaviour that we have, and so we should take them into account when judging the likelihood of democratic failure. For instance, wealthier and more educated people are much less likely to support populist or extremist groups. But Runciman focuses the most on age, which I think is the correct approach. Wealth is relative - even if people are actually much richer, they can feel poor and angry after a recession (as they did in the 1930s, despite still being many times wealthier than almost all their ancestors). Education may just be correlated with other factors, rather than the actual cause of lasting differences in mindset. But there are pretty clear biological and social reasons to think that the behaviour and priorities of older people are robustly and significantly different from those of younger people. You need only look at the age distribution of violent crime, for example, to see how strong this effect is (although it may have lessened somewhat over recent years, since marriage rates are declining and single men cause more trouble).

In short: the failures of democracy in the 30's were based on large populations of young men who could be mobilised in anger by militaristic leaders - see for instance the brownshirts in Germany and blackshirts in Italy. But that's not what the failure of democracy in our time would look like, because that group of people is much smaller now. For better or worse, older populations are less disruptive and more complacent. To see where that might lead, consider Japan: an ageing population which can't drag itself out of economic torpor, resistant to immigration, dominated for decades by one political party, betting the country's future on using robots to replace the missing workforce.

Changes ahead

During a Q&A after the talk, I pointed out that Japan is very different to western countries in its particularly strong culture of social conformity and stability. Age trends notwithstanding, I have much more difficulty imaging the same quiet tolerance of slow decline occurring in the US or UK. So, given that government institutions are very difficult to change, where will people direct their frustration if lacklustre growth continues in the coming decades?

In response, Runciman raised two possibilities. Firstly, that people will "go around their governments", finding new domains in which politics is less relevant. We could call this the "Wild West" possibility. Of course, there's no longer an uncolonised West to explore - but there is the internet, which isn't democratically run and probably never will be. We already see fewer young men working full-time, because the alternative of spending most of their time gaming has become more appealing. As virtual worlds become even more immersive, it seems plausible that people will begin to care much less about political issues.

One problem with the idea of "going around governments", though, is that governments are just much bigger now than they used to be. And as technology companies profit from the growing role of the internet, there'll be pressure for governments to intervene even more to fight inequality. So a second option is a more Chinese approach, with increasingly autocratic Western governments exerting heavy pressure on (and perhaps eventually control over) tech companies.

A more optimistic possibility is for the internet to make democracy more accountable. Runciman invites us to consider Plato's original argument against direct democracy (in which people vote on individual issues) - that it would lead to rule by the poor, the

ignorant, and the young, all of whom necessarily outnumber the wealthy, wise and old. This argument turned out not to apply for representative democracy, since elected representatives tend to be wealthy, educated and old despite their constituents being the opposite. But now it's inapplicable for a different reason - that although our representatives haven't changed much, the rest of us are starting to look much more like them. So maybe it'll become feasible to implement a more direct democracy, facilitated by the internet and modern communication technology. (This still seems like a bad idea to me, though.)

Base rates and complacency

The last section was a little speculative, so let's take a step back and think about how to make predictions about these sorts of events in general. Runciman's analysis above provides good reasons not to draw a specific parallel between the rise of fascism last century and recent political events. But it would take extraordinary evidence to exempt us from extrapolating broader historical trends, in particular the fact that states always collapse eventually, and that the base rate for coups and other forms of internal strife is fairly high. Are the extraordinary changes we've seen since the industrial revolution sufficient to justify belief in our exceptionalism?

It's true that since World War 2, almost no wealthy democracies have descended into autocracy or chaos (Turkey and Ireland being two edge cases). It's also true that, despite widespread political disillusionment, norms against violence have held to a remarkably large degree. But drawing judgements from the historical period "since World War 2" is a classic case of the Texan Sharpshooter's Fallacy (and possibly also anthropic bias?). In fact, this recent lull should make us skeptical about our ability to evaluate the question objectively, because people are in general very bad at anticipating extreme events that haven't occurred in living memory. I think this is true despite these possibilities being discussed in the media. For example, while there's a lot of talk about Trump being a potential autocrat, few Americans are responding by stockpiling food or investing in foreign currencies or emigrating. This suggests that hostility towards Trump is driven primarily by partisan politics, rather than genuine concern about democratic collapse. An additional data point in favour of this hypothesis is how easily the Republican political establishment has fallen in line.

Another key question which isn't often discussed is the nature of modern military culture. Historically, this has been a major factor affecting governmental stability. But, apart from vague intuitions about modern militaries being fairly placid, I find myself remarkably ignorant on this subject, and suspect others are as well. What facts do you know about your country's military, about the character of its commanders or the distribution of power within it, that make you confident that it won't launch a coup if, for example, one of its generals is narrowly defeated in a disputed presidential election (as in Gore vs Bush)? Note that military demographics haven't changed nearly as much as those of our societies overall. They're still primarily composed of young working-class men without degrees - a group that's unusually angry about today's politics. So while I am pretty convinced by Runciman's arguments, this is one way in which they may not apply. Also consider that warfare is much less hands-on than it used to be, and firepower much more centrally concentrated, both of which make coups easier.

And what about extreme events?

So far I've looked at societal collapse from a political point of view. But many historical transitions were precipitated by natural disasters or diseases. See, for instance, the Mayan collapse, or the Little Ice Age, or the Black Death. Today, we're much safer from natural disasters, both because of our technology and because of the scale of our societies - few people live in countries in which the majority of a population can be struck by a single natural disaster. Similarly, we're also much safer from natural diseases. But we're much more vulnerable to severe man-made disasters, which I think are very likely to occur over the next century. Since this post is focused on political collapse as a distinct phenomenon to technological disaster, I won't discuss extreme risks from technology here. However, it's worthwhile to look at the ways in which smaller technological harms might exacerbate other trends. AI-caused unemployment and the more general trend towards bimodal outcomes in western countries are likely to cause social unrest. Meanwhile terrorism is going to become much easier - consider being able to 3D-print assassin drones running facial recognition software, for instance. And due to antibiotic overuse, it's likely that our safety from disease will decline over the coming years (even without the additional danger of bioterrorism using engineered diseases). Finally, I think we're much softer than we used to be - it won't take nearly as much danger to disrupt a country. Runciman is probably correct that we're less susceptible to a collapse into authoritarianism than we were in the past - but the same trends driving that change are also pushing us towards new reasons to worry.

In addition to the talk by Runciman, this post was inspired by discussions with my friends Todor and Julio, and benefited from their feedback.

Embedded Curiosities

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

A final word on curiosity, and intellectual puzzles:

I described an embedded agent, Emmy, and said that I don't understand how she evaluates her options, models the world, models herself, or decomposes and solves problems.

In the past, when researchers have talked about motivations for working on problems like these, they've generally focused on the motivation from [AI risk](#). AI researchers want to build machines that can solve problems in the general-purpose fashion of a human, and [dualism](#) is not a realistic framework for thinking about such systems. In particular, it's an approximation that's especially prone to breaking down as AI systems get smarter. When people figure out how to build general AI systems, we want those researchers to be in a better position to understand their systems, analyze their internal properties, and be confident in their future behavior.

This is the motivation for most researchers today who are working on things like updateless decision theory and subsystem alignment. We care about basic conceptual puzzles which we think we need to figure out in order to achieve confidence in future AI systems, and not have to rely quite so much on brute-force search or trial and error.

But the arguments for why we may or may not need particular conceptual insights in AI are pretty long. I haven't tried to wade into the details of that debate here. Instead, I've been discussing a particular set of research directions as an *intellectual puzzle*, and not as an instrumental strategy.

One downside of discussing these problems as instrumental strategies is that it can lead to some misunderstandings about *why* we think this kind of work is so important. With the "instrumental strategies" lens, it's tempting to draw a direct line from a given research problem to a given safety concern. But it's not that I'm imagining real-world embedded systems being "too Bayesian" and this somehow causing problems, if we don't figure out what's wrong with current models of rational agency. It's certainly not that I'm imagining future AI systems being written in second-order logic! In most cases, I'm not trying at all to draw direct lines between research problems and [specific AI failure modes](#).

What I'm instead thinking about is this: We sure do seem to be working with the wrong basic concepts today when we try to think about what agency is, as seen by the fact that these concepts don't transfer well to the more realistic embedded framework.

If AI developers in the future are *still* working with these confused and incomplete basic concepts as they try to actually build powerful real-world optimizers, that seems like a bad position to be in. And it seems like the research community is unlikely to figure most of this out by default in the course of just trying to develop more capable systems. Evolution certainly figured out how to build human brains without "understanding" any of this, via brute-force search.

Embedded agency is my way of trying to point at what I think is a very important and central place where I feel confused, and where I think future researchers risk running into confusions too.

There's also a lot of excellent AI alignment research that's being done with an eye toward more direct applications; but I think of that safety research as having a different type signature than the puzzles I've talked about here.

Intellectual curiosity isn't the ultimate reason we privilege these research directions. But there are some *practical* advantages to orienting toward research questions from a place of curiosity at times, as opposed to *only applying the "practical impact" lens* to how we think about the world.

When we apply the curiosity lens to the world, we orient toward the sources of confusion preventing us from seeing clearly; the blank spots in our map, the flaws in our lens. It encourages re-checking assumptions and attending to blind spots, which is helpful as a psychological counterpoint to our "instrumental strategy" lens—the latter being more vulnerable to the urge to lean on whatever shaky premises we have on hand so we can get to more solidity and closure in our early thinking.

Embedded agency is an organizing theme behind most, if not all, of our big curiosities. It seems like a central mystery underlying many concrete difficulties.

The Ubiquitous Converse Lawvere Problem

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

(This post was originally published on Oct 20th 2017, and is 1 of 4 posts brought forwarded today as part of the AI Alignment Forum launch sequence on fixed points.)

In this post, I give a stronger version of the open question presented [here](#), and give a motivation for this stronger property. This came out of conversations with Marcello, Sam, and Tsvi.

Definition: A continuous function $f : X \rightarrow Y$ is called ubiquitous if for every continuous function $g : X \rightarrow Y$, there exists a point $x \in X$ such that $f(x) = g(x)$.

Open Problem: Does there exist a topological space X with a ubiquitous function $f : X \rightarrow [0, 1]^X$?

I will refer to the [original](#) problem as the Converse Lawvere Problem, and the new version as the the Ubiquitous Converse Lawvere Problem. I will refer to a space satisfying the conditions of (Ubiquitous) Converse Lawvere Problem, a (Ubiquitous) Converse Lawvere Space, abbreviated (U)CLS. Note that a UCLS is also a CLS, since a ubiquitous is always surjective, since g can be any constant function.

Motivation: True FairBot

Let X be a Converse Lawvere Space. Note that since such an X might not exist, the following claims might be vacuous. Let $f : X \rightarrow [0, 1]^X$ be a continuous surjection.

We will view X as a space of possible agents in an open source prisoner's dilemma game. Given two agents $A, B \in X$, we will interpret $f_A(B)$ as the probability with which A cooperates when playing against B . We will define $U_A(B) := 2f_B(A) - f_A(B)$, and interpret this as the utility of agent A when playing in the prisoner's dilemma with B .

Since f is surjective, every continuous policy is implemented by some agent. In particular, this means gives:

Claim: For any agent $A \in X$, there exists another agent $A' \in X$ such that $f_{A'}(B) = f_B(A)$.

i.e. A' responds to B the way that B responds to A .

Proof: The function $B \mapsto f_B(A)$ is a continuous function, since $B \mapsto f_B$ is continuous, and evaluation is continuous. Thus, there is a policy $B \mapsto f_B(A)$ in $[0, 1]^X$. Since f is surjective, this policy must be the image under f of some agent A' , so $f_{A'}(B) = f_B(A)$.

Thus, for any fixed agent A , we have some other agent A' that responds to any B the way B responds to A . However, it would be nice if $A' = A$, to create a FairBot that responds to any opponent the way that that opponent responds to it. Unfortunately, to construct such a FairBot, we need the extra assumption that f is ubiquitous.

Claim: If f is ubiquitous, then exists a true fair bot in X : an agent $FB \in X$, such that $f_{FB}(A) = f_A(FB)$ for all agents $A \in X$.

Proof: Given an agent $B \in X$, there exists an policy $g_B \in [0, 1]^X$ such that $g_B(A) = f_A(B)$ for all A , since $A \mapsto f_A(B)$ is continuous. Further, the function $B \mapsto g_B$ is continuous, since the function $A, B \mapsto f_A(B)$ and the definition of the exponential topology. Since f is ubiquitous, there must be some $FB \in X$ such that $f_{FB} = g_{FB}$. But then, for all A , we have $f_{FB}(A) = g_{FB}(A) = f_A(FB)$.

Note that we may not need the full power of ubiquitous here, but it is the simplest property I see that gets the result.

Note that this FairBot is fair in a stronger sense than the FairBot of [modal combat](#), in that it always has the same output as its opponent. This may make you suspicious, since the you can also construct an UnfairBot, UB such that $f_{UB}(A) = 1 - f_A(UB)$ for all

A . This would have caused a problem in the modal combat framework, since you can put a FairBot and an UnfairBot together to form a paradox. However, we do not have this problem, since we deal with probabilities, and simply have

$f_{UB}(FB) = f_{FB}(UB) = 1/2$. Note that the exact phenomenon that allows this to possibly work is the fixed point property of the interval $[0, 1]$ which is the only reason that we cannot use diagonalization to show that no CLS exists.

Finally, note that we already have a combat framework that has a true FairBot: the [reflective oracle](#) framework. In fact, the reflective oracle framework may have all the benefits we would hope to get out of a UCLS. (other than the benefit of simplicity of not having to deal with computability and hemicontinuity).

This post was originally published on Oct 20th 2017, and has been brought forwarded as part of the AI Alignment Forum launch sequences.

Tomorrow's AIAF sequences post will be 'Iterated Amplification and Distillation' by Ajeya Cotra, in the sequence on iterated amplification.

Implementations of immortality

This is a linkpost for <http://thinkingcomplete.blogspot.com/2018/04/implementations-of-immortality.html>

(Inspired by [Eliezer's essays on designing utopias.](#))

I was recently talking to a friend about what key features society would need for people to be happy throughout arbitrarily-long lives - in other words, what would a utopia for immortals look like? He said that it would require continual novelty. But I think that by itself, novelty is under-specified. Almost every game of Go ever played is novel and unique, but eventually playing Go would get boring. Then you could try chess, I suppose - but at a certain point the whole concept of board games would become tiresome. I don't think bouncing around between many types of different activities would be much better, in the long run. Rather, the sort of novelty that's most desirable is a change of perspective, such that you find meaning in things you didn't appreciate before. That interpretation of novelty is actually fairly similar to my answer; I said that the most important requirement is a feeling of progress. By this I mean:

- Your past isn't being lost as it recedes from you.
- Your future will be better than your past - in qualitative ways as well as quantitative.
- You receive increasing social recognition for your achievements.
- You feel like you are continually growing as a person.

Some of these criteria are amenable to technical solutions - for example, memory enhancements would be very helpful for the first. But simply abiding by the basic principles of liberalism makes others very difficult to universalise. As long as we allow people to make their own choices, there will be some people who end up falling into addiction, or lethargy, or self-destructive spirals. We could theoretically make this rarer by having stronger social or legal norms, so that people still have freedom, but not total freedom. We could also make the consequences of failure less unpleasant (e.g. with welfare systems) and less permanent (e.g. by eradicating the most addictive drugs). Yet even then, the social repercussions of being unsuccessful would still weigh on people heavily - man does not live by bread alone, but by the status judgements of his peers.

Fortunately, perceived social status is not zero-sum, because different subcultures value different things. That helps many more people receive social recognition for their achievements; ideally everyone would find [a Dunbar-sized community](#), in which they can distinguish themselves. Sure, some will be envious of other communities, but I think abstract concerns like those would mostly be outweighed by their tangible activities and relationships. For example (although I don't have good data on this) anecdotally it seems that modern homeless communities can often be tighter-knit and more supportive than well-off suburbs.

But splintering society into fragments doesn't solve the question of what the overarching cultural framework should look like - and we do need one. Firstly because subcultures usually need something to define themselves in opposition to; also because moving between subcultures would be much more difficult if they didn't share fundamental tenets. Yet now we're back to the problem of what general society

should prize and reward in order for almost everyone to feel like their lives are valuable and progressing towards even more value.

Here's a slightly unusual solution, which embraces Eliezer's recommendation that utopias should be weird: we should strictly stratify society by age. This doesn't mean that people of different ages must isolate themselves from each other (although some will), but rather that:

- Older people are respected greatly simply by virtue of their age.
- Access to some prestigious communities or social groups is age-restricted.
- There are strong norms (or even rules?) about which types of activities one should do at a given age.

Age hierarchies are not a new idea; they've been the norm throughout human history, and were only relatively recently discarded from western culture. Granted, that was for good reasons, like the fact that they tend to hold back social and technological progress. But our challenge here is to come up with a steady-state culture which can provide lasting happiness, not one which maximises speedy progress. Age hierarchies are the one sort of hierarchy in which everyone gets to advance arbitrarily far upwards. They also tap into fundamental aspects of human motivation. Video games are addictive because you can keep unlocking new content or "levelling up". But they're also frustrating because that progress isn't grounded in anything except a counter on the screen. Games like Starcraft and League of Legends instead provide satisfaction through victory over others - but that's zero-sum, which we don't want. The third class of games which people spend most time on - MMORPGs such as World of Warcraft or EVE - augment the experience of gaining resources and levelling up with a community in which high-achieving players are respected. In a long-lived society with age hierarchies, people could always look forward to "unlocking new content" based on their age. Even people who aren't very high-status within their own age group could find respect amongst younger groups; and as long as people continued having children, your relative status in society would always increase.

Note that this doesn't imply that children and young adults would have bad lives. Firstly, despite being respected the least, they also experience the greatest sense of novelty and opportunity. Even if they feel like they're missing out on some opportunities, they can be consoled by the thought that their time will come. And they probably won't be very upset about "missing out" anyway - the experiences which older people prize highly are often not those which young people envy. Teens don't feel like their lives are worse because they don't yet have children and go to nightclubs not cocktail parties. Champagne-sipping parents, on the other hand, usually think their lives have become deeper and richer since their teenage years; neither side is unhappy with their lot. A more speculative example which comes to mind is the traditional progression through stages of enlightenment in Buddhism, where you simply don't understand what you're missing out on until you're no longer missing it. To take this to an extreme, access to the next stage of society could depend on learning a new language or framework of knowledge, in which you can discuss ideas you couldn't even conceptualise before.

Stepping back from the weirder implementations, how might longevity impact close relationships? There's a story I remember reading about a world in which people live arbitrarily long; every few decades, though, they simply leave their entire social circle, move away, and start afresh. This seems like a reasonable way to keep a sense of excitement and novelty alive. However, my friend argues that if you live for a very long time, and become close to many people, then you'll eventually stop thinking of

them as unique, valuable individuals. I do think that there are enough possible ways to have relationships, and enough variation between people, to last many, many lifetimes, so that you can continually be surprised and grateful, and develop as a person. But I concede that left to themselves, people wouldn't necessarily seek out this variety - they might just decide on a type, and stick to it. I think that age stratification helps solve this problem too. Perhaps there can be expectations for how to conduct relationships based on your age bracket: at some points cultivating a few deep friendships, at others being a social butterfly; sometimes monogamous, sometimes polyamorous; sometimes dating people similar to you, sometimes people totally different; sometimes staying within your age group, and sometimes spending time with people who are much older or younger and have totally different perspectives. In our world, people would shirk from this - but I imagine that a very long-lived society would have a culture more open to trying new things.

Lastly, we should remember that the dark side of having strong social norms is enforcement of those norms. To some extent this can be avoided by creating a mythos which people buy into. Cultural narratives affect people on such a deep level that many never question the core tenets (like, in the west, the value of individualism). It would also help if there were separate communities which people could join as an act of rebellion, instead of wreaking havoc in their original one. But in general, creating norms such that even people who challenge those norms do so in non-destructive ways seems like a very difficult problem. We're basically trying to find stable, low-entropy configurations for a chaotic system (as opposed to stable high-entropy configurations, such as total collapse). Even worse, the system is self-referential - individuals within it can reason about the system as a whole, and some will then try to subvert it. There's much more which needs to be figured out, including entirely new fields of research - but nobody ever said designing a utopia would be easy.

On MIRI's new research directions

This is a linkpost for <https://intelligence.org/2018-update-our-new-research-directions/>

Nate Soares has written up a post that discusses MIRI's new research directions: a mix of reasons why we're pursuing them, reasons why we're pursuing them the way we are, and high-level comparisons to Agent Foundations.

Read the full (long!) post [here](#).

Double-Dipping in Dunning--Kruger

Originally posted at sandymaguire.me.

I have to get you to drop modesty and say to yourself, "Yes, I would like to do first-class work." Our society frowns on people who set out to do really good work. You're not supposed to; luck is supposed to descend on you and you do great things by chance. Well, that's a kind of dumb thing to say. I say, why shouldn't you set out to do something significant. You don't have to tell other people, but shouldn't you say to yourself, "Yes, I would like to do something significant."

Richard Hamming

I want to talk about impostor syndrome today. If you're unfamiliar with the term, it's this phenomenon where people discount their achievements and feel like frauds---like at any moment, others will realize that they don't belong. People with impostor syndrome are convinced that they've somehow pulled the wool over others' eyes, and have conned their way into their jobs or social circles. And it's very, very common.

On the other end of the spectrum, there's another phenomenon known as Dunning--Kruger. This one is widely reported in the media as "the average person believes they are better than 70% of drivers." They can't all be right---by definition, 50% of people must be worse than the median driver. This too is very common; it seems to hold true in most positive-feeling, every-day domains.

While this is indeed supported by Dunning--Kruger's data, [the paper itself](#) focuses on the fact that "the worst performers in a task are often *so bad that they have no idea*." For example, the least funny people are so tragically unfunny that they wouldn't know humor if it bit them. As a result, unfunny people are *entirely blind* to their lack of humor:

Participants scoring in the bottom quartile on our humor test not only overestimated their percentile ranking, but they overestimated it by 46 percentile points.

A less well-known finding of Dunning--Kruger is that the best performers will systematically *underestimate* how good they are, by about 15 percentile points. The proposed reason is that they found the task to be easy, assume others must have also found it easy, and go from there. In other words, top performers are *so good that they don't notice many challenges*.

It's unfortunate that Dunning--Kruger has been popularized as "most people think they are better than they are." Not only do high-performers already underestimate themselves, but those who *know about Dunning--Kruger in its popularized form* are likely to double-dip, and *further adjust down to compensate for this*. For example, if you are in fact in the top 90% for some skill, due to Dunning--Kruger you will likely estimate yourself to be at 75%. *But*, if you know that people routinely overestimate their abilities by 20%, you might drop your estimate down to 55% in compensation---significantly lower than your true skill level.

If this is true, it would suggest some of the world's best people will estimate themselves to be *worse* than the self-evaluations of the worlds' worst. The takeaway is this: if you're the kind of person who worries about statistics and Dunning--Kruger in

the first place, you're already *way above average* and clearly have the necessary meta-cognition to not fall victim to such things. From now on, unless you have evidence that you're *particularly bad at something*, I want you to assume that you're 15 percentile points higher than you would otherwise estimate.

The mathematician Richard Hamming is said to have often ruffled feathers by asking his colleagues "what's the most important problem in your field, and why aren't you working on it?" Most people, I'd suspect, would say that the most important problems are too hard to be tackled by the likes of them. That it would take minds greater minds than theirs to make progress on such things. They give up before having even tried.

Instead, the smartest people I know join Google or go work at B2B startups and are simultaneously bored in their day jobs, feel like they're frauds while they're there, and don't have enough energy to work on personal projects after hours. But at least they're making wicked-big paychecks for themselves. And for their less-qualified leadership.

The best minds of my generation are thinking about how to make people click ads.

Jeff Hammerbacher

Here's an interesting thought experiment. If you randomly swapped places with someone for a week---you doing their job and they doing yours---how would it turn out? If you're a competent programmer with passable social skills, I suspect you would be a lot more successful in that week than your replacement. Most things *just aren't that hard*, and a good percentage of the people doing those jobs are phoning it in anyways.

The bar on competency is tragically low. And yet the world revolves.

If you agree with this line of reasoning, it means the world is just *oozing* with potential, ready for the taking. Most of the world's most competent people are unaware of *just how good they are*. Most things really aren't as hard as getting through that graph-theory class you took, and don't take nearly as much effort. The world is being run by people who are too incompetent to know it; people who are only in power because they're the ones who showed up, and because showing up is most of the battle.

Which leads us to the inescapable conclusion that this world we live in is particularly amenable to change. That if you're willing to trust in your own instinct and tackle hard-seeming problems, you're going to experience literally unbelievable amounts of success. Everyone else is deferring to better minds. So be the change you want to see in the world, because we've been waiting for you.

Is Copenhagen or Many Worlds true? An experiment. What? Yes.

This is a linkpost for <https://arxiv.org/abs/1811.02983v1>

If this paper is right, an experiment should be possible.

Building it also doesn't seem that hard to do technically, so I think it could be interesting to watch in the coming months/years.

If Many Worlds dies, we would seem to have a perpetual motion machine of the second kind. But if the second law of thermodynamics prevails, Copenhagen should provably be wrong.

All, if this paper is right. Does somebody find a mistake?

New safety research agenda: scalable agent alignment via reward modeling

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

This is a linkpost for <https://medium.com/@deepmindsafetyresearch/scalable-agent-alignment-via-reward-modeling-bf4ab06dfd84>

Jan Leike and others from the DeepMind safety team have released a new [research agenda](#) on reward learning:

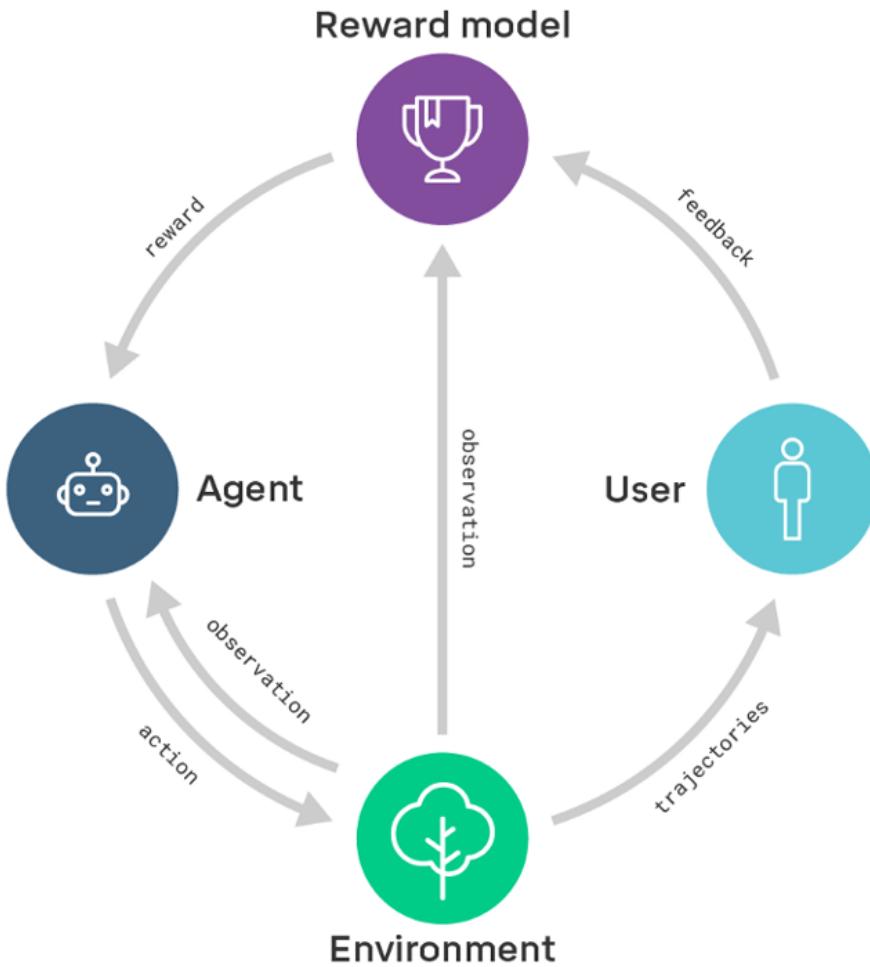
"Ultimately, the goal of AI progress is to benefit humans by enabling us to address increasingly complex challenges in the real world. But **the real world does not come with built-in reward functions**. This presents some challenges because performance on these tasks is not easily defined. We need a good way to provide feedback and enable artificial agents to reliably understand what we want, in order to help us achieve it. In other words, we want to train AI systems with human feedback in such a way that the system's behavior *aligns* with our intentions. For our purposes, we define the **agent alignment problem** as follows:

How can we create agents that behave in accordance with the user's intentions?

The alignment problem can be framed in the reinforcement learning framework, except that instead of receiving a numeric *reward signal*, the agent can interact with the user via an interaction protocol that allows the user to communicate their intention to the agent. This protocol can take many forms: the user can provide [demonstrations](#), [preferences](#), [optimal actions](#), or [communicate a reward function](#), for example. **A solution to the agent alignment problem is a policy that behaves in accordance with the user's intentions.**

With our [new paper](#) we outline a research direction for tackling the agent alignment problem head-on. Building on our earlier [categorization of AI safety problems](#) as well as [numerous problem expositions on AI safety](#), we paint a coherent picture of how progress in these areas could yield a solution to the agent alignment problem. This opens the door to building systems that can better understand how to interact with users, learn from their feedback, and predict their preferences—both in narrow, simpler domains in the near term, and also more complex and abstract domains that require understanding beyond human level in the longer term.

The main thrust of our research direction is based on **reward modeling**: we train a reward model with feedback from the user to capture their intentions. At the same time, we train a policy with reinforcement learning to maximize the reward from the reward model. In other words, we **separate** learning **what** to do (the reward model) from learning **how** to do it (the policy).



For example, in previous work we taught agents to [do a backflip from user preferences](#), to [arrange objects into shapes with goal state examples](#), to [play Atari games from user preferences and expert demonstrations](#). In the future we want to design algorithms that learn to adapt to the way users provide feedback (e.g. using natural language)."

If You Want to Win, Stop Conceding

Author's note: This is the first in what I suppose might be a series of posts with respect to things I learned from playing traditional games competitively that I think might have broader applications. I write this as a private individual, not on behalf of CFAR or any other organization.

Traditional games - card games, board games, miniatures, etc. are a lot of fun, and I've played several of them at quite a competitive level. [1]

The #1 piece of advice that I can give if you want to get better at these games - a piece of advice that applies across essentially every game or sport I've played and a lot of "real world" stuff as well - is "*if you want to win, stop conceding.*"

On the surface that doesn't sound super deep or interesting, but there's more to it than the obvious meaning - not all concessions are formal resignations, and indeed the ones that aren't are often more important.

Some time ago I read a book - either "The Inner Game of Tennis" or "Bonds that Make Us Free" or maybe both - that taught me that very many people concede games well before they need to be over, either because they incorrectly estimate their chances or because they make a mental motion away from trying to win and towards trying to make excuses for losing.

Here are some examples of what excuse-making thoughts might sound like:

- "The dice are against me, there's nothing I can do."
- "I didn't get a good night's sleep or eat breakfast this morning, otherwise I would be winning - I just can't focus."
- "This guy's bad but he got lucky." [2]
- "This guy's way better than me, I shouldn't even be matched up against him." [3]
- "I don't know how she even got this far ahead, there must be some bug."
- "This is pointless, why play it out?"
- "This is just for fun anyway, I've basically lost, may as well wind things down."

Once you have moved from trying to win and towards trying to excuse losing, you have more or less already lost. Sometimes an opponent might snatch defeat from the jaws of victory, but that's rare. Most of the time, assuming you've lost is the same as actually losing, it just takes longer.

If you want to get better, *stop it*. Force yourself to keep playing and keep looking for outs. Learn to recognize these mental patterns and suppress them. Don't concede games unless there's nothing you can do anymore. Fight until the bitter end.

That's a real 'if', because if you don't want to get better I certainly don't recommend doing this! There is a sense in which it is viscerally unpleasant to force yourself to be in a losing position, desperately scrabbling for anything that can get you out.

When you get in that position and escape, it feels great - but there are also going to be times when you don't escape, you struggle for ten or fifteen or thirty minutes but

still lose, the whole thing feels bad, and you might wish you had conceded in the first place. If you aren't prepared to face that, maybe don't bother.

But what I've found is that a practice of facing the negative thoughts and pushing through seems to have been quite beneficial to me across a wide range of games and areas, so I would give serious thought to the notion that - at least in areas that you care about and want to be better at - you should consider this approach.

A few days before writing this, I played a card game where around two to five times throughout the course of an ~hour-long game I was struck by thoughts along the lines of "Wow, my position is horrible. I've been really unlucky. I should concede."

I didn't concede, and I won the game. This is not a particularly unusual experience to have once you acquire the inclination and ability to push through.

[1] For calibration:

At various times I have been world #1 by Elo rating at a few different games I've played (not all at the same time!). I came in third at the World Championships of L5R this year despite being out of practice (though to be fair I had some good luck).

I have flown to a card game tournament in another state in large part because I calculated I was likely to win enough in prizes that I would net gain money from the trip; I haven't eBayed all the promotional items I won yet but I believe I made hundreds of dollars in expectation from that venture.

I don't say this to boast but rather to give an indication of where I am coming from. In order to "deflate the sails" a bit, I should say that I do not make a career out of gaming and I don't play poker or Magic: the Gathering competitively, which have a significantly higher [level of play](#) than most games I play; there are many people who are better at games than I am.

[2] This thought pattern is especially bad because it prevents you from learning from the game after the fact. Sure, some games do come down to luck in the end, but probably there were decisions you could have made prior to that that would influence the odds.

[3] A joke saying goes: "Anyone worse than me at this game is casual n00b trash. Anyone better than me at this game is a no-life tryhard." Neither of the thoughts in that dichotomy is very useful to have, even in their less straw forms.

On Rigorous Error Handling

In my long career as a programmer I've seen a lot of projects with a lot of error handling practices and, sadly, almost none that actually worked.

We are facing few high-level problems here:

1. For many applications error handling is not critical.
2. Error handling code is rarely executed. Consequently, it looks like it works until it doesn't.
3. Programmers want to implement new features. Writing error handling is just an annoyance that slows them down.

Who cares about error handling?

For many applications writing rigorous error handling code doesn't pay off.

As long as you have some semi-decent way to deal with errors transparently (e.g. exceptions) you are fine. If there's a problem, throw an exception. On the top level catch all the exceptions and open a dialog box with the error message (in frontend applications) or write it to the log (on the server).

The above, of course, means that you hope that, after the exception is processed, the application will be left in a consistent, functional state. Once again, you HOPE. You do not KNOW. You do not know because, when writing the application, you haven't thought about error handling at all.

Now, don't get me wrong. I am not saying that you should never do this. It often makes sense from economic perspective. Writing rigorous error handling is expensive. If the worst thing that can happen is that a frontend application crashes once in a long while making user curse and restart it, then it's totally not worth doing.

However, there is also a different kind of application. There are medical application where the patient dies if they misbehave. There are space probes that crash into planets, taking millions of dollars of investment as well as scientific careers down with them. There are HFT application which can generate bazillion of dollars of loss for every minute of downtime.

In these cases it's definitely worth writing rigorous error handling code.

That much everybody agrees with.

However, it is often not explicitly understood that whenever you are writing a piece of infrastructure (as opposed to a client-facing application) you are always in the latter category.

A piece of infrastructure can be used in a medical application, in a space probe, in a high-frequency trading system, but unless it's a commercial product and you take care not to sell it to anyone who looks too important to fail, then you have no way to prevent that.

Ergo, if you are doing infrastructure, you should always do rigorous error handling.

Error handling code is never executed

The mainstream attitude to fixing bugs at the moment is "run it, see what fails, then fix it". That's why we write tests. Tests make the code run and, hopefully, fail if there's a problem. That's why we have user-facing issue trackers. If a bug gets past testing, it will fail for the user and user will fill in a bug report — or so we hope, at least.

But the error handling code is, almost by definition, executed only rarely. And code that is triggered by two concurrent problems is executed almost never at all.

Therefore, tests are not likely to help. User reports may catch the issue, but are often closed with "cannot reproduce" status.

There are alternative approaches, like formal verification or fuzzing, but almost nobody uses those.

In the end, we want programmers writing error handling code that works on the first try, without any testing.

And that is, of course, impossible to do. But, on the other hand, it's not black and white. Errors are rare and if the bugs in error handling code get rare as well then you can increase the reliability of the system from, say, three nines to five nines. And as medical applications go, adding one nine of reliability saves some actual human lives.

The dilemma

So here we are. We have programmers who want to spend no time on error handling, who just want to move on and implement cool new features, and we want them to write perfect error handling code on the first try, without testing.

Even worse, attempts to solve the first problem (to make error handling invisible) often make the second problem (failure-proof error handling) even more complex. And given that everyone is writing new code all the time, but errors are rare, the solutions to the first problem are going to proliferate even at the expense of making the second problem harder or even impossible to solve.

An example

Let's have a look at OpenSSL API. I don't want to pick on OpenSSL, mind you. You'll encounter similar kind of problem with almost any library, but I happened to work with OpenSSL lately so that's what I'll use as an example. Also note that OpenSSL is notoriously underfunded, so instead of getting angry at them, do your bit and send them a donation.

Anyway, when an OpenSSL function fails it reports that it failed and you can have a look at the "error queue". Error queue is effectively a list of integer error codes. The list of error codes can be found [here](#).

Note how documentation of an error code invariably looks like this:

```
#define ERR_R_BAD_GET ASN1_OBJECT_CALL 60 Definition at line 296 of file err.h.
```

Yes, you get a cryptic string, a cryptic number, and a reference to the line in the source code. If you look at the source code you'll find a cryptic string and a cryptic number again. No explanation of what the error means whatsoever.

Now think for a second about how you would, as a user, handle such an error. It's a multidimensional beast and it's not even clear whether order of errors matters or not. And each error is itself chosen from a large set of cryptic errors with no explicit associated semantics at all. Needless to say, the set of error codes is going to expand in newer versions, so even if you handled every single one of them correctly you can still get a nasty failure at some point in the future.

So what I've ended up doing was logging all the errors from the queue for debugging purposes and converting the entire thing into a single "OpenSSL failed" error. I am not proud of that but I challenge you to come up with a better solution.

And no, passing the monstrosity to the caller, who has even less context about how to handle it than you have, and breaking the encapsulation properties along the way, is not an option.

The takeaway from the example is that even if you are willing to do rigorous error handling, the underlying layers often give you no chance but to go the easy and sloppy way.

As a side note, OpenSSL's system of reporting errors is by no means the worst of all. Consider typical exceptions as they appear in most languages. Exception is not a list of error codes. It's an object. In other words, it can be a list, a map, a binary tree or whatever you want, really. There's absolutely no way to deal with this Turing-complete beast in a systemic way.

What's needed

When you look at it from the point of view of the user of the API, the requirements are suddenly crystal-clear. If you are to do rigorous error handling you want a small amount of well-defined failure modes.

A function may succeed. Or it may, for example, fail because of disconnected backend. Or it may time out. And that's it. There are only two failure modes and they are documented as a part of the API. Once you have that, the error handling becomes obvious:

```
int err = some_library_function(); if(err != 0) { switch(err) { case ECONNRESET: ...  
case ETIMEDOUT: ... default: assert(0); // the function violates its specification } }
```

Some libraries are closer to this ideal state, some are further away.

As already mentioned, classic exceptions are the worst. You get an object, or worse, just an interface that you can (the horror!) downcast to an actual exception class. All bets are off.

OpenSSL's approach is somewhat better. You still get a list of error codes, but at least you know it's a list and not a map or something else.

Some libraries go further in the right direction and return a single error, but the list of error codes is long, poorly documented and subject to the future growth. Moreover,

there's no clear association between a function and the error codes it can return. You have to expect any error code when calling any function. Often, the rigorous error handling code, as show above, would have to have dozens if not hundreds of error codes in the switch statement.

When possible errors are part of the function specification, on the other hand, we are almost OK. This is the case with [I've also tried to used the same approach with my own libraries, such as \[http://libdill.org/documentation.html libdill\]](#). However, even in this case it's possible to do it wrong. The list of possible errors in POSIX specification of [view-source:<https://pubs.opengroup.org/onlinepubs/009695399/functions/connect.html>] lists more than a dozen of possible errors which in turn disincentivizes the user from doing rigorous error handling.

In the best of the cases, the list of errors is small not only for a particular function, but also for the library as a whole. In my experience, it's entirely possible to do with as little as 10-15 different error codes for the entire library.

The advantage of that approach is twofold.

First, the user, as they use the library, will eventually learn those error codes and what they mean by heart. That in turn makes writing rigorous error handling code much less painful.

Second, a limited set of error codes makes the implementer of the library to actually think about the failure modes. If he can just define a new error code for every error condition he encounters, it's an easy way out with no need to do much thinking. If, on the other hand, he has to choose one of 10 possible error code, he has to think about which of them matches the semantics of the error the best. Which in turn results in a better, more consistent library, which it is joy to use.

November 17th, 2018

Clarifying "AI Alignment"

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

When I say an AI A is *aligned with* an operator H, I mean:

A is trying to do what H wants it to do.

The “alignment problem” is the problem of building powerful AI systems that are aligned with their operators.

This is significantly narrower than some other definitions of the alignment problem, so it seems important to clarify what I mean.

In particular, this is the problem of getting your AI to try to do the right thing, **not** the problem of figuring out which thing is right. An aligned AI would try to figure out which thing is right, and like a human it may or may not succeed.

Analogy

Consider a human assistant who is trying their hardest to do what H wants.

I’d say this assistant is aligned with H. If we build an AI that has an analogous relationship to H, then I’d say we’ve solved the alignment problem.

“Aligned” doesn’t mean “perfect:”

- They could misunderstand an instruction, or be wrong about what H wants at a particular moment in time.
- They may not know everything about the world, and so fail to recognize that an action has a particular bad side effect.
- They may not know everything about H’s preferences, and so fail to recognize that a particular side effect is bad.
- They may build an unaligned AI (while attempting to build an aligned AI).

I use alignment as a statement about the *motives* of the assistant, not about their knowledge or ability. Improving their knowledge or ability will make them a better assistant—for example, an assistant who knows everything there is to know about H is less likely to be mistaken about what H wants—but it won’t make them *more aligned*.

(For very low capabilities it becomes hard to talk about alignment. For example, if the assistant can’t recognize or communicate with H, it may not be meaningful to ask whether they are aligned with H.)

Clarifications

- The definition is intended [*de dicto* rather than *de re*](#). An aligned A is trying to “do what H wants it to do.” Suppose A thinks that H likes apples, and so goes to the store to buy some apples, but H really prefers oranges. I’d call this behavior

aligned because A is trying to do what H wants, even though the thing it is trying to do (“buy apples”) turns out not to be what H wants: the *de re* interpretation is false but the *de dicto* interpretation is true.

- An aligned AI can make errors, including moral or psychological errors, and fixing those errors isn’t part of my definition of alignment except insofar as it’s part of getting the AI to “try to do what H wants” *de dicto*. This is a critical difference between my definition and some other common definitions. I think that using a broader definition (or the *de re* reading) would also be defensible, but I like it less because it includes many subproblems that I think (a) are much less urgent, (b) are likely to involve totally different techniques than the urgent part of alignment.
- An aligned AI would also be trying to do what H wants **with respect to clarifying H’s preferences**. For example, it should decide whether to ask if H prefers apples or oranges, based on its best guesses about how important the decision is to H, how confident it is in its current guess, how annoying it would be to ask, etc. Of course, it may also make a mistake at the meta level—for example, it may not understand when it is OK to interrupt H, and therefore avoid asking questions that it would have been better to ask.
- This definition of “alignment” is extremely imprecise. I expect it to correspond to some more precise concept that cleaves reality at the joints. But that might not become clear, one way or the other, until we’ve made significant progress.
- One reason the definition is imprecise is that it’s unclear how to apply the concepts of “intention,” “incentive,” or “motive” to an AI system. One naive approach would be to equate the incentives of an ML system with the objective it was optimized for, but this seems to be a mistake. For example, humans are optimized for reproductive fitness, but it is wrong to say that a human is incentivized to maximize reproductive fitness.
- “What H wants” is even more problematic than “trying.” Clarifying what this expression means, and how to operationalize it in a way that could be used to inform an AI’s behavior, is part of the alignment problem. Without additional clarity on this concept, we will not be able to build an AI that tries to do what H wants it to do.

Postscript on terminological history

I [originally](#) described this problem as part of “the AI control problem,” following Nick Bostrom’s usage in *Superintelligence*, and used “the alignment problem” to mean “understanding how to build AI systems that share human preferences/values” (which would include efforts to clarify human preferences/values).

I adopted the new terminology after some people expressed concern with “the control problem.” There is also a slight difference in meaning: the control problem is about coping with the possibility that an AI would have different preferences from its operator. Alignment is a particular approach to that problem, namely avoiding the preference divergence altogether (so excluding techniques like “put the AI in a really secure box so it can’t cause any trouble”). There currently seems to be a tentative consensus in favor of this approach to the control problem.

I don’t have a strong view about whether “alignment” should refer to this problem or to something different. I do think that *some* term needs to refer to this problem, to separate it from other problems like “understanding what humans want,” “solving philosophy,” etc.

This post was originally published [here](#) on 7th April 2018.

The next post in this sequence will post on Saturday, and will be "An Unaligned Benchmark" by Paul Christiano.

Tomorrow's AI Alignment Sequences post will be the first in a short new sequence of technical exercises from Scott Garrabrant.

The Vulnerable World Hypothesis (by Bostrom)

This is a linkpost for <https://nickbostrom.com/papers/vulnerable.pdf>

Nick Bostrom has put up a new working paper to his personal site (for the first time in two years?), called *The Vulnerable World Hypothesis*.

I don't think I have time to read it all, but I'd be interested to see people **comment with some choice quotes** from the paper, and also read **people's opinions on the ideas within it**.

To get the basics, below I've written down the headings into a table of contents, copied in a few definitions I found when skimming, and also copied over the conclusion (which seemed to me more readable and useful than the abstract).

Contents

- Is there a black ball in the urn of possible inventions?
- A thought experiment: easy nukes
- The vulnerable world hypothesis
 - 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.
- Typology of vulnerabilities
 - Type-1 ("easy nukes")
 - Type-1 vulnerability: There is some technology which is so destructive and so easy to use that, given the semi-anarchic default condition, the actions of actors in the apocalyptic residual make civilizational devastation extremely likely.
 - Type-2a ("safe first strike")
 - Type-2a vulnerability: There is some level of technology at which powerful actors have the ability to produce civilization-devastating harms and, in the semi-anarchic default condition, face incentives to use that ability.
 - Type-2b ("worse global warming")
 - Type-2b vulnerability: There is some level of technology at which, in the semi-anarchic default condition, a great many actors face incentives to take some slightly damaging action such that the combined effect of those actions is civilizational devastation.
 - Type-0 ("surprising strangelets")
 - Type-0 vulnerability: There is some technology that carries a hidden risk such that the default outcome when it is discovered is inadvertent civilizational devastation. 47
- Achieving stabilization
 - Technological relinquishment
 - Principle of Differential Technological Development. Retard the development of dangerous and harmful technologies, especially ones that raise the level of existential risk; and accelerate the

- development of beneficial technologies, especially those that reduce the existential risks posed by nature or by other technologies.
- Preference modification
- Some specific countermeasures and their limitations
- Governance gaps
- Preventive policing
- Global governance
- Discussion
- Conclusion

Conclusion

This paper has introduced a perspective from which we can more easily see how civilization is vulnerable to certain types of possible outcomes of our technological creativity—our drawing a metaphorical black ball from the urn of inventions, which we have the power to extract but not to put back in. We developed a typology of such potential vulnerabilities, and showed how some of them result from destruction becoming too easy, others from pernicious changes in the incentives facing a few powerful state actors or a large number of weak actors.

We also examined a variety of possible responses and their limitations. We traced the root cause of our civilizational exposure to two structural properties of the contemporary world order: on the one hand, the lack of preventive policing capacity to block, with extremely high reliability, individuals or small groups from carrying out actions that are highly illegal; and, on the other hand, the lack of global governance capacity to reliably solve the gravest international coordination problems even when vital national interests by default incentivize states to defect. General stabilization against potential civilizational vulnerabilities—in a world where technological innovation is occurring rapidly along a wide frontier, and in which there are large numbers of actors with a diverse set of human-recognizable motivations—would require that both of these governance gaps be eliminated. Until such a time as this is accomplished, humanity will remain vulnerable to drawing a technological black ball.

Clearly, these reflections prove a pro tanto reason to support strengthening surveillance capabilities and preventive policing systems and for favoring a global governance regime that is capable of decisive action (whether based on unilateral hegemonic strength or powerful multilateral institutions). However, we have not settled whether these things would be desirable all-things-considered, since doing so would require analyzing a number of other strong considerations that lie outside the scope of this paper.

Because our main goal has been to put some signposts up in the macrostrategic landscape, we have focused our discussion at a fairly abstract level, developing concepts that can help us orient ourselves (with respect to long-term outcomes and global desirabilities) somewhat independently of the details of our varying local contexts.

In practice, were one to undertake an effort to stabilize our civilization against potential black balls, one might find it prudent to focus initially on partial solutions and low-hanging fruits. Thus, rather than directly trying to bring about extremely effective preventive policing or strong global governance, one might attempt to patch up particular domains where black balls seem most likely to appear. One

could, for example, strengthen oversight of biotechnology-related activities by developing better ways to track key materials and equipment, and to monitor activities within labs. One could also tighten know-your-customer regulations in the biotech supply sector, and expand the use of background checks for personnel working in certain kinds of labs or involved with certain kinds of experiment. One can improve whistleblower systems, and try to raise biosecurity standards globally. One could also pursue differential technological development, for instance by strengthening the biological weapons convention and maintaining the global taboo on biological weapons. Funding bodies and ethical approval committees could be encouraged to take broader view of the potential consequences of particular lines of work, focusing not only on risks to lab workers, test animals, and human research subjects, but also on ways that the hoped-for findings might lower the competence bar for bioterrorists down the road. Work that is predominantly protective (such as disease outbreak monitoring, public health capacity building, improvement of air filtration devices) could be differentially promoted.

Nevertheless, while pursuing such limited objectives, one should bear in mind that the protection they would offer covers only special subsets of scenarios, and might be temporary. If one finds oneself in a position to influence the macroparameters of preventive policing capacity or global governance capacity, one should consider that fundamental changes in those domains may be the only way to achieve a general ability to stabilize our civilization against emerging technological vulnerabilities.

Stabilize-Reflect-Execute

You've recently joined a major organization in a senior management role. How can you organize your plans?

One simple way to think about them is with what can be called the "Stabilize-Reflect-Execute" cycle.

Stabilize

You first check if there are any urgent issues and address them immediately. Are there burning problems or opportunities that need to be dealt with? Second, you do anything you need to do to best prepare yourself for reflection. If there are people you need to talk to in order to get necessary information, you set that up upfront.

Reflect

Once urgent issues are dealt with and you are able to properly access the situation, you work to do so. For executives this can mean a lengthy period of discussions with all of the relevant people and thoughts on strategy before making formal announcements. This could take a few weeks or months.

Execute

Now is the time to begin working on non-urgent important problems, which should be the main ones. You follow through with your reflection. Execution may involve deciding on pursuing future larger stabilize-reflect-execute loops.

Let's summarize. "*Stabilize*" refers to handling urgent issues and preparing for reflection. This is similar to the notion of getting one's "[house in order](#)." "*Reflect*" refers to deciding how to best deal with the important non-urgent issues. "*Execute*" refers to working on the important issues. This is basically a subset of the Eisenhower Method for situations where these three steps make up the majority of the work.

Examples

I think this cycle plays out in many important situations, so may be worth some independent study. Some examples of these cycles include:

	Stabilize	Reflect	Execute
Personal career selection	If you have serious mental, physical, monetary, or family problems, try to deal with them.	Evaluate your career options.	Begin a new line of work or educational program.
Getting things done	If there are any particularly urgent & important tasks (the house is literally on fire), do those first. Make sure you have adequate time & space for the reflection stage. Get coffee.	Go through the first steps of the GTD process to clarity, organize, and prioritize your work.	Do the work.
Setting up a research organization	Obtain access to funding and initial talent.	Write a research agenda and experiment with initial research efforts.	Follow through on a research agenda and deliver the work in useful ways.
Initiating political leadership	Deal with immediate crises. Ensure access to funds. Ensure that no immediate rebellions will occur.	Review all confidential information. Work with key advisors to modify your plans to account for new information.	Carry out strategy.
An ideal global governance structure	Get our "house in order." Make sure that there aren't any pressing threats to humanity or the ability to do a significant reflection.	Engage in a long session of consideration and debate on how to best shape the future.	Shape the future.

Necessary Conditions

The stabilize-reflect-execute cycle is good for some specific situations. I think it may require the following:

1. There are some tasks that are both urgent and important.

If this is not true, the "stabilize" step isn't necessary.

2. There are some tasks that are both non-urgent and important.

If this is not true, the "reflect" and "execute" steps aren't necessary.

3. There is significant expected value for spending time figuring out how to do the non-urgent and important tasks.

If this is not true, the "reflect" step isn't necessary.

4. Figuring out how to do the important tasks can be done in one batch per full cycle.

If this is not true, then the work may be a bit of a mess of stabilizations, reflections, and executions, as opposed to one serial process.

Implications

We could use this model to compare use cases

I think that different use cases (like the examples above) share a good amount of similarities. I hope that they could be seen as such, and used to help understand each other. One may think that we don't have many similar references classes to things like AI takeoffs and global governance. While this may be generally true, I hope that this similarity could provide a bit of a counter.

We should acknowledge uncertainty of the “Execute” stage

One trying to understand the stabilize-reflect-execute of an actor may try to predict the actions of the execution stage, but I think this should be regarded with skepticism. The point of the reflect stage is to better understand how to enact the execute stage; its' presence suggests that the execution stage could go in different ways. This means that a lot of attention on stabilize-reflect-execute processes should be on the first two parts, rather than the third.

Comparison to the Eisenhower Method

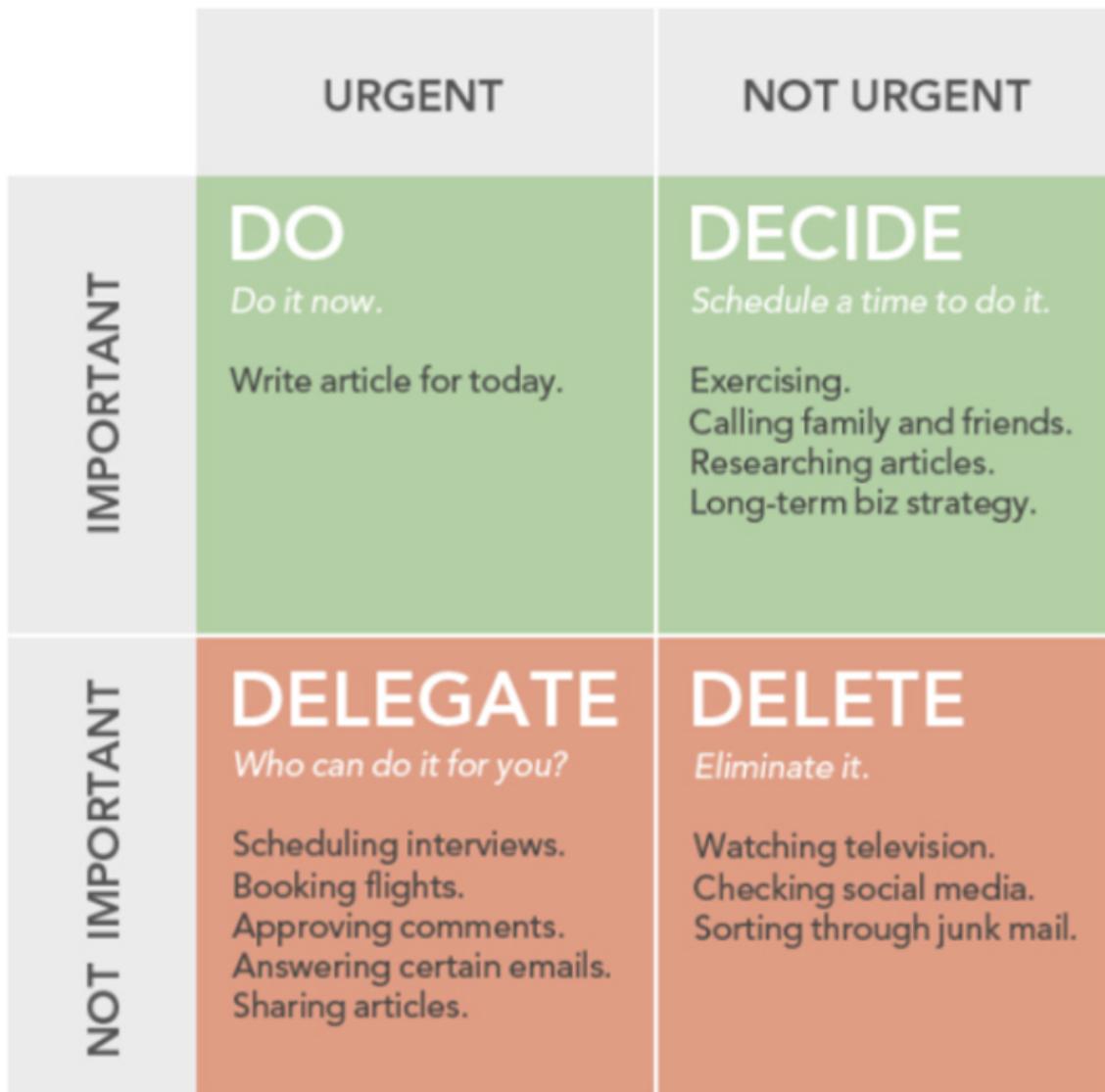


Image from [James Clear](#)

The urgent / important distinction comes from the [Eisenhower Matrix](#). That matrix is supposed to apply to more situations, so is more generic. Work around the matrix typically recommends "quickly doing" the "important & urgent" tasks, then "planning" the "important & non urgent" tasks.

Comparison to Decision Cycles

Wikipedia lists several examples of interesting [decision cycles](#). These are sequences with specific steps for decision making. For instance, in quality control, "PDCA (Plan-Do-Check-Act) is used." I believe the decision cycles typically include some kind of learning component, which is absent in the stabilize-reflect-execute cycle.

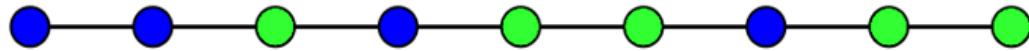
Many thanks to Toby Ord for discussion & inspiration, and to Owen Cotton-Barratt and Max Daniel for feedback.

Topological Fixed Point Exercises

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

This is one of three sets of fixed point exercises. The first post in this sequence is [here](#), giving context.

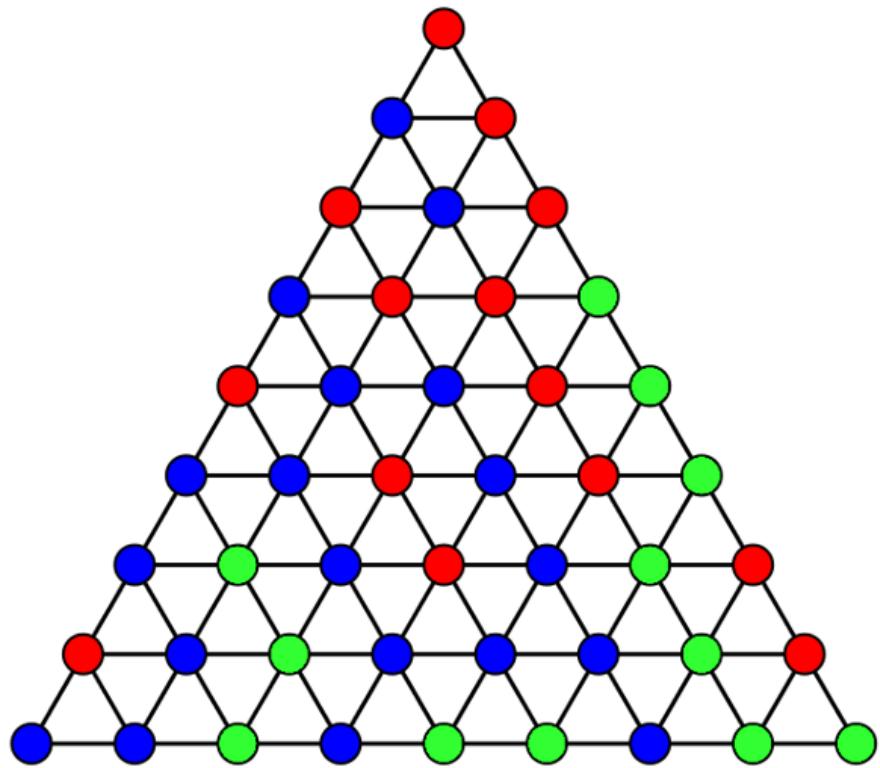
- 1.** (1-D Sperner's lemma) Consider a path built out of n edges as shown. Color each vertex blue or green such that the leftmost vertex is blue and the rightmost vertex is green. Show that an odd number of the edges will be bichromatic.



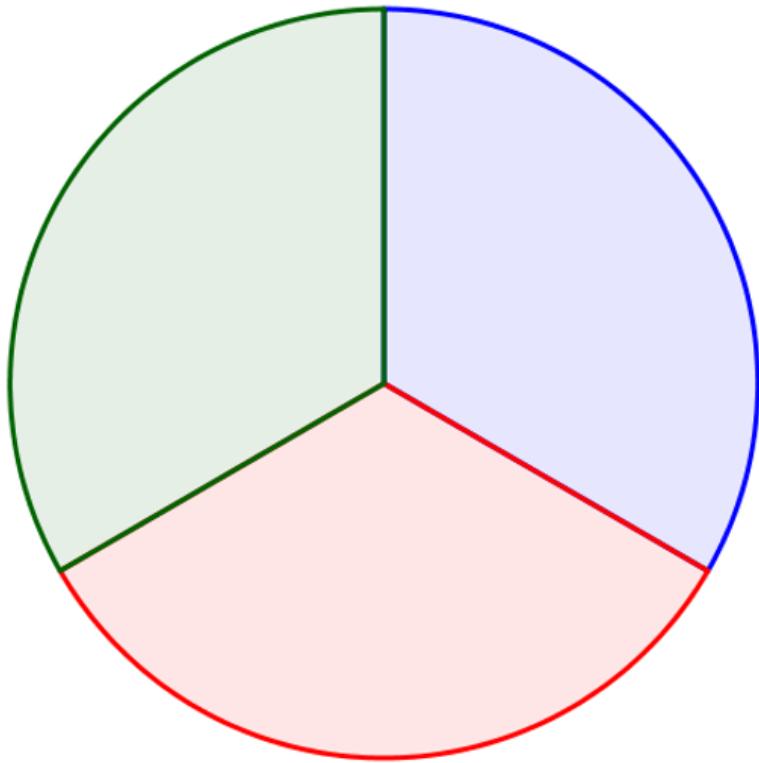
- 2.** (Intermediate value theorem) The Bolzano-Weierstrass theorem states that any bounded sequence in \mathbb{R}^n has a convergent subsequence. The intermediate value theorem states that if you have a continuous function $f : [0, 1] \rightarrow \mathbb{R}$ such that $f(0) \leq 0$ and $f(1) \geq 0$, then there exists an $x \in [0, 1]$ such that $f(x) = 0$. Prove the intermediate value theorem. It may be helpful later on if your proof uses 1-D Sperner's lemma and the Bolzano-Weierstrass theorem

- 3.** (1-D Brouwer fixed point theorem) Show that any continuous function $f : [0, 1] \rightarrow [0, 1]$ has a fixed point (i.e. a point $x \in [0, 1]$ with $f(x) = x$). Why is this not true for the open interval $(0, 1)$?

- 4.** (2-D Sperner's lemma) Consider a triangle built out of n^2 smaller triangles as shown. Color each vertex red, blue, or green, such that none of the vertices on the large bottom edge are red, none of the vertices on the large left edge are green, and none of the vertices on the large right edge are blue. Show that an odd number of the small triangles will be trichromatic.



5. Color the all the points in the disk as shown. Let f be a continuous function from a closed triangle to the disk, such that the bottom edge is sent to non-red points, the left edge is sent to non-green points, and the right edge is sent to non-blue points. Show that f sends some point in the triangle to the center.



- 6.** Show that any continuous function f from closed triangle to itself has a fixed point.
- 7.** (2-D Brouwer fixed point theorem) Show that any continuous function from a compact convex subset of \mathbb{R}^2 to itself has a fixed point. (You may use the fact that given any closed convex subset S of \mathbb{R}^n , the function from \mathbb{R}^n to S which projects each point to the nearest point in S is well defined and continuous.)
- 8.** Reflect on how non-constructive all of the above fixed-point findings are. Find a parameterized class of functions where for each $t \in [0, 1]$, $f_t : [0, 1] \rightarrow [0, 1]$, and the function $t \mapsto f_t$ is continuous, but there is no continuous way to pick out a single fixed point from each function (i.e. no continuous function g such that $g(t)$ is a fixed point of f_t for all t).
- 9.** (Sperner's lemma) Generalize exercises 1 and 4 to an arbitrary dimension simplex.
- 10.** (Brouwer fixed point theorem) Show that any continuous function from a compact convex subset of \mathbb{R}^n to itself has a fixed point.

11. Given two nonempty compact subsets $A, B \subseteq R^n$, the Hausdorff distance between them is the supremum

$$\max \{ \sup_{a \in A} d(a, B), \sup_{b \in B} d(b, A) \}$$

over all points in either subset of the distance from that point to the other subset. We call a set valued function $f : S \rightarrow 2^T$ a continuous Hausdorff limit if there is a sequence $\{f_n\}$ of continuous functions from S to T whose graphs, $\{(x, y) \mid y = f_n(x)\} \subseteq S \times T$, converge to the graph of f , $\{(x, y) \mid f(x) \ni y\} \subseteq S \times T$, in Hausdorff distance. Show that every continuous Hausdorff limit $f : T \rightarrow 2^T$ from a compact convex subset of R^n to itself has a fixed point (a point x such that $x \in f(x)$).

12. Let S and T be nonempty compact convex subsets of R^n . We say that a set valued function, $f : S \rightarrow 2^T$ is a Kakutani function if the graph of f , $\{(x, y) \mid f(x) \ni y\} \subseteq S \times T$, is closed, and $f(x)$ is convex and nonempty for all $x \in S$. For example, we could take S and T to be the interval $[0, 1]$, and we could have $f : S \rightarrow 2^T$ send each $x < \frac{1}{2}$ to $\{0\}$, map $x = \frac{1}{2}$ to the whole interval $[0, 1]$, and map $x > \frac{1}{2}$ to $\{1\}$. Show that every Kakutani function is a continuous Hausdorff limit. (Hint: Start with the case where S is a unit cube. Construct f_n by breaking S into small cubes of side length 2^{-n} . Construct a smaller cube of side length 2^{-n-1} within each 2^{-n} cube. Send each small 2^{-n-1} to the convex hull of the images of all points in the 2^{-n} cube with a continuous function, and glue these together with straight lines. Make sure you don't accidentally get extra limit points.)

13. (Kakutani fixed point theorem) Show that every Kakutani function from a compact convex subset of $S \subseteq R^n$ to itself has a fixed point.

Please use the spoilers feature - the symbol '>' followed by '!' followed by space -in your comments to hide all solutions, partial solutions, and other discussions of the math. The comments will be moderated strictly to hide spoilers!

I recommend putting all the object level points in spoilers and leaving metadata outside of the spoilers, like so: "I think I've solved problem #5, here's my solution <spoilers>" or "I'd like help with problem #3, here's what I understand <spoilers>" so that people can choose what to read.

Fixed Point Discussion

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

Warning: This post contains some important spoilers for [Topological Fixed Point Exercises](#), [Diagonalization Fixed Point Exercises](#), and [Iteration Fixed Point Exercises](#). If you plan to even try the exercises, reading this post will significantly reduce the value you can get from doing them.

Core Ideas

A [fixed point](#) of a function f is an input x such that $f(x) = x$. Fixed point theorems show that various types of functions must have fixed points, and sometimes give methods for finding those fixed points.

Fixed point theorems come in three flavors: Topological, Diagonal, and Iterative. (I sometimes refer to them by central examples as Brouwer, Lawvere, and Banach, respectively.)

Topological fixed points are non constructive. If f is continuous, $f(0) > 0$, and $f(1) < 1$, then we know f must have some fixed point between 0 and 1, since $f(x)$ must somewhere transition from being greater than x to being less than x . This does not tell us where it happens. This can be especially troublesome when there are multiple fixed points, and there is no principled way to choose between them.

Diagonal fixed points are constructed with a weird trick where you feed a code for a function into that function itself. Given a function f , if you can construct a function g , which on input x , interprets x as a function, runs x on itself, and then runs f on the result (i.e. $g(x) := f(x(x))$), then $g(g)$ is a fixed point of f because $g(g) = f(g(g))$. This is not just an example; everything in the cluster looks like this. It is a weird trick, but it is actually very important.

Iterative fixed points can be found through iteration. For example, if $f(x) = -x$, then starting with any x value, iterating f forever will converge to the unique fixed point $x = 0$.

(There is a fourth cluster in number theory discussed [here](#), but I am leaving it out, since it does not seem relevant to AI, and because I am not sure whether to put it by itself or to tack it onto the topological cluster.)

Topological Fixed Points

Examples of topological fixed point theorems include [Sperner's lemma](#), [the Brouwer fixed point theorem](#), [the Kakutani fixed point theorem](#), [the intermediate value theorem](#), and [the Poincaré-Miranda theorem](#).

- Brouwer is the central example of the cluster. Brouwer states that any continuous function f from a compact convex set to itself has a fixed point.
- Sperner's Lemma is a discrete analogue which is used in one proof of Brouwer.
- Kakutani is a strengthening of Brouwer to degenerate set valued functions, that look almost like continuous functions.
- Poincaré-Miranda is an alternate formulation of Brouwer, which is about finding zeros rather than fixed points.
- The Intermediate Value Theorem is a special case of Poincaré-Miranda. To a first approximation, you can think of all of these theorems as one big theorem.

Topological fixed point theorems also have some very large applications. The Kakutani fixed point theorem is used in game theory [to show that Nash equilibria exist](#), and [to show that markets have equilibrium prices!](#) Sperner's lemma is also used in [some envy-free fair division results](#). Brouwer is also used to show the existence of [some differential equations](#).

In MIRI's agent foundations work, Kakutani is used to construct [probabilistic truth predicates](#) and [reflective oracles](#), and Brouwer is used to construct [logical inductors](#).

These applications all use topological fixed points very directly, and so carry with them most of the philosophical baggage of topological fixed points. For example, while Nash equilibria exist, they are not unique, are computationally hard to find, and feel non-constructive and arbitrary.

Diagonal Fixed Points

Diagonal fixed point theorems are all centered around the basic structure of $g(g)$,

where $g(x) := f(x(x))$, as mentioned previously.

The pattern is used in many places.

- In CS theory, it is used to construct [quines](#) and the [Y-combinator in lambda calculus](#), and to prove [Rice's theorem](#) and that the halting problem is undecidable.
- In formal logic, it is used to prove the diagonal lemma and important corollaries, like [Gödel's incompleteness theorem](#), [Löb's theorem](#), and [Tarski's undefinability theorem](#). It is used to show the uncountability of the real numbers with [Cantor's Diagonal Argument](#).
- [Lawvere's fixed point theorem](#) is the most general version of the argument, and can be used to show all of the above as a corollary.

In MIRI's agent foundations work, this shows up in the [Löbian obstacle to self-trust](#), [Löbian handshakes in Modal Combat](#) and [Bounded Open Source Prisoner's Dilemma](#),

as well as providing a basic foundation for why an agent reasoning about itself might make sense at all through quines.

Iterative Fixed Points

Iterative fixed point theorems are less of one cluster than the others; I will factor it into two sub-clusters, centered around the [Banach fixed point theorem](#) and [Tarski fixed point theorem](#). (Each the same size as the original.)

The Tarski cluster is about fixed points of monotonic functions on (partially) ordered sets found by iteration. Tarski's fixed point theorem states that any order preserving function on a complete lattice has a fixed point (and further the set of fixed points forms a complete lattice). The least fixed point can be found by starting with the least element and iterating the function transfinitely. This, for example, implies that every monotonic function on from $[0, 1]$ to itself has a fixed point, even if it is not continuous. [Kleene's fixed point theorem](#) strengthens the assumptions of Tarski by adding a form of continuity (and also removes some irrelevant assumptions), which gives us that the least fixed point can be found by iterating the function only ω times. [The fixed point lemma for normal functions](#) is similar to Kleene, but with ordinals rather than partial orders. It states that any strictly increasing continuous function on ordinals has arbitrarily large fixed points.

The Banach cluster is about fixed points of contractive functions on metric spaces found by iteration. A contractive function is a function that sends points closer together. A function f is contractive if there exists an $\epsilon > 0$ such that for all $x \neq y$ $d(f(x), f(y)) \leq (1 - \epsilon)d(x, y)$. Banach's fixed point theorem state that any contractive function has a unique fixed point. This fixed point is $\lim_{n \rightarrow \infty} f^n(x)$ for any starting point x . An application of this to linear functions is that any ergodic stationary Markov chain has a stationary distribution (which is a fixed point of the transition map), which is converged to via iteration. This is also used in showing that correlated equilibria exist and can be found quickly. Banach can also be used to show that gradient descent converges exponentially quickly on a strongly convex function.

Interdisciplinary Nature

I think of Pure Mathematics as divided at the top into 5 subfields: [Algebra](#), [Analysis](#), [Topology](#), [Logic](#), and [Combinatorics](#).

The mapping of the key fixed point theorems discussed in the exercises into these categories is surjective:

- Lawvere's fixed point theorem is Algebra
- Banach's fixed point theorem is Analysis
- Brouwer fixed point theorem is Topology
- Gödel's first incompleteness theorem is Logic
- Sperner's lemma is Combinatorics.

On top of that, major applications of fixed point theorems show up in Differential Equations, CS theory, Machine Learning, Game Theory, and Economics.

Tomorrow's AI Alignment Forum sequences post will be two short posts 'Approval-directed bootstrapping' and 'Humans consulting HCH' by Paul Christiano in the sequence Iterated Amplification.

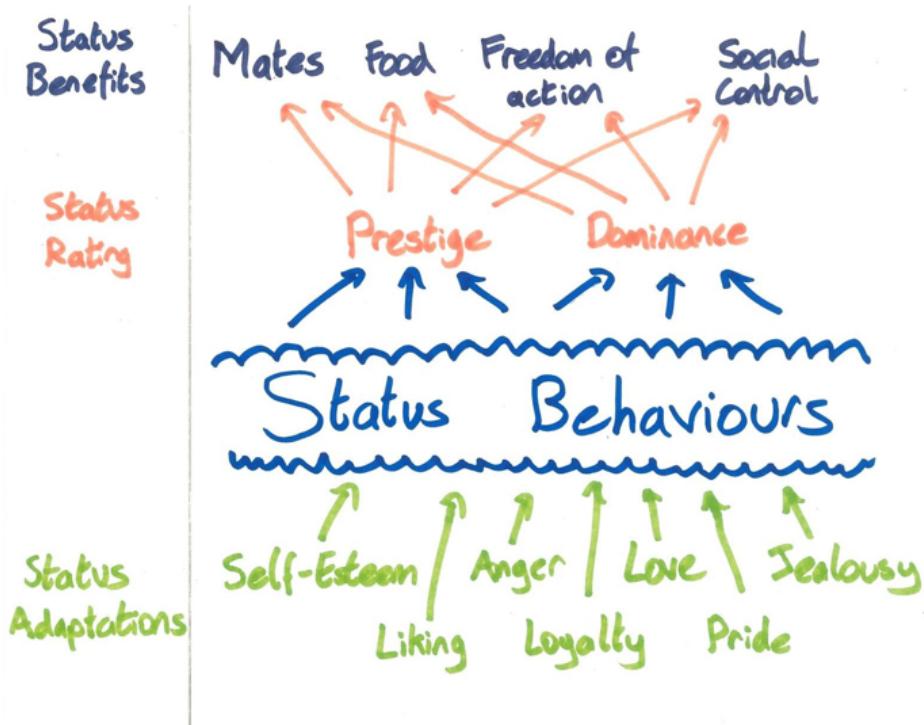
The next posting in this sequence will be four posts of Agent Foundations research that use fixed point theorems, on Wednesday 28th November. These will be re-posts of content from the now-defunct Agent Foundations forum, all of whose content is now findable on the AI Alignment Forum (and all old links will soon be re-directed to the AI Alignment Forum).

Status model

[Epistemic status: My best guess.]

Following [a conversation](#) on a previous post, I decided to do some research into the community's thoughts on status. The community talks about status a lot and I spent a few happy hours sifting through everything I could find.

I came up with a status model based on the posts and comments which I read:



Essentially, status-related mental adaptations are executed which lead to certain status behaviours. These behaviours determine our social standing which, at least in the ancestral environment, tended to affect our fitness.

I don't think there's anything groundbreaking here (a similar model would probably apply to any adaptation execution vs fitness maximising effect) but I haven't seen it sketched out specifically for status before.

The word "status" gets used to refer to items on all 4 levels and this can lead to confusion where two people are referring to different levels.

For instance, what is status and how can I measure it? One way is to look directly at row 2 (status rating): who respects whom and how much? Maybe I could give everyone a questionnaire to rate each other's status. Or I can look at row 3 (status behaviours): who acts like they have high status? Or row 1 (status benefits): who has the most social control etc.? Whoever gets the most benefits probably has the most status.

Each option has its advantages and disadvantages e.g. accuracy, ease of assessment but it is important to know which level is being referred to.

Probably the most commonly debated issue on this topic is whether status is zero-sum.

If we consider status in row 2 then status is probably going to be relatively zero-sum, although you can maybe get around this a bit by splitting into smaller sub-cultures.

If we consider row 1, insofar as the status rating determines the benefits, they are close to zero-sum. However, the benefits are also controlled by things other than status (how good are we at getting food, how well coordinated are we as a group?) and so are able to be positive sum.

Row 4 is where it gets really interesting - our adaptations which implement status behaviour are not zero-sum. We can feel more self-esteem without increasing our status rating (see [That Other Kind Of Status](#)). These mental adaptations are the things which we care about on a gut level and give plenty of scope for positive sum behaviours (e.g. [give praise](#)).

I don't pretend that this answers the zero-sum question completely but I think it does put it in a helpful frame.

The model is incomplete in a number of ways.

The "status rating" row is a massive simplification. In reality there are all of the different ways which humans judge status, how status changes depending on group and circumstances and the effects of social allies. I only listed [two kinds of status](#) to simplify visually.

The status benefits and status adaptations listed are also only a subset of the actual benefits and adaptations.

The relationships between the rows are leaky. The status adaptations lead to behaviours which aren't necessarily related to status and the status benefits can be affected by things other than status rating.

Despite the model's limitations, I hope it is a useful simplification.

EDIT:

Ben Pace [suggested](#) I add a list of the posts which I looked at and how I think they relate to my model. I've done this below - apologies to any of the contributors if I'm misinterpreting their work.

[That Other Kind Of Status](#): Link between status rating and status adaptations.

[The Many Faces of Status](#): Morendil's own investigation into status. Covers all areas of the model.

[The Red Paperclip Theory of Status](#): Trading within and between status ratings and status benefits plus some discussion of related behaviours.

[Status: Is it what we think it is?](#): Discussion of prestige and dominance status.
[Diegocaleiro](#) gives the academic names of the status types and links to google scholar for more info.

[Actors see status](#): Quotes from Impro, focusing on status behaviour and status rating.

[The Economics of Social Status](#): Inspired by Red Paperclip and goes into significant detail.

[Status: Map and Territory](#): Discussion of how some status adaptations attempt to enforce the integrity of the links between other adaptations and status behaviours.

[Making yourself small](#): How status rating plays into status benefits, specifically freedom of action. Low status rating can make it hard to make yourself big. High status rating allows you to make yourself big or small.

[What if status is a terminal value for most people](#): This seems to me to be a different model, where status rating is a direct adaptation. Some commenters (e.g. [someonewrongonthenet](#)) suggest that separating status from adaptation works better.

[The conversation which started my research](#) – To predict status behaviours, should we go backwards from status rating (zero-sum) or is there a better way (e.g. working forwards from status adaptations)?

Speculative Evopsych, Ep. 1

(cw death, religion, suicide, evolutionary psychology, shameless tongue-in-cheek meta-contrarianism)

I have a passing interest in biology, so I recently bought some fruit flies to run experiments on. I did two things to them. First, I bred them for intelligence. The details are kinda boring, so let's fast-forward: after a few tens of millions of generations, they were respectably intelligent, with language and culture and technology so on.

In parallel with that, and more interestingly, whenever a fly was about to die of injury, I immediately plucked it out of the box, healed it, and put it in a *different* box ("Box Two"), a magnificent paradise where it blissfully lived out the rest of its days. Evolutionarily, of course, relocation was equivalent to death, and so the flies evolved to treat them the same: you could still make somebody stop affecting your world by stabbing them, and their kin would still grieve and seek revenge – the only difference was the lack of a corpse.

It didn't really matter that the two boxes were separated only by a pane of glass, and that the flies in Box One could clearly see their "deceased" fellows living fantastic lives in Box Two. They "knew" on an abstract, intellectual level that getting "fatally" wounded wouldn't actually make them stop having conscious experiences like death would. But *evolution doesn't care about that distinction*; so it doesn't select for organisms that care about that distinction; so the flies generally disregarded Box Two.

A small subculture in Box One claimed that "if anybody *actually* believed in Box Two and all its wonders, they'd stab themselves through the heart in order to get there faster. Everybody's literally-mortal fear of relocation proves that they don't *truly* believe in Box Two, they only – at best – *believe they believe*."

Strangely, nobody found this argument convincing.

Believing others' priors

Meet the Bayesians

In one way of looking at Bayesian reasoners, there are a bunch of possible worlds and a bunch of people, who start out with some guesses about what possible world we're in. Everyone knows everyone else's initial guesses. As evidence comes in, agents change their guesses about which world they're in via Bayesian updating.

The Bayesians can share information just by sharing how their beliefs have changed.

"Bob initially thought that last Monday would be sunny with probability 0.8, but now he thinks it was sunny with probability 0.9, so he must have seen evidence that he judges as 4/9ths as likely if it wasn't sunny than if it was"

If they have the same priors, they'll converge to the same beliefs. But if they don't, it seems they can agree to disagree. This is a bit frustrating, because we don't want people to ignore our *very convincing evidence* just because they've gotten away with having a stupid weird prior.

What can we say about which priors are permissible? Robin Hanson offers an argument that we must either (a) believe our prior was created by a special process that correlated it with the truth more than everyone else's or (b) our prior must be the same as everyone else's.

Meet the pre-Bayesians

How does that argument go? Roughly, Hanson describes a slightly more nuanced set of reasoners: the pre-Bayesians. The pre-Bayesians are not only uncertain about what world they're in, but also about what everyone's priors are.

These uncertainties can be tangled together (the joint distribution doesn't have to factorise into their beliefs about everyone's priors and their beliefs about worlds). Facts about the world can change their opinions about what prior assignments people have.

Hanson then imposes a pre-rationality condition: if you find out what priors everyone has, you should agree with your prior about how likely different worlds are. In other words, you should trust your prior in the future. Once you have this condition, it seems that it's impossible to both (a) believe that some other people's priors were generated in a way that makes them as likely to be good as yours and (b) have different priors from those people.

Let's dig into the sort of things this pre-rationality condition commits you to.

Consider the class of worlds where you are generated by a machine that randomly generates a prior and sticks it in your head. The pre-rationality rule says that worlds where this randomly-generated prior describes the world well are more likely than worlds where it is a poor description.

So if I pop out with a very certain belief that I have eleven toes, such that no amount of visual evidence that I have ten toes can shake my faith, the pre-prior should indeed place more weight on those worlds where I have eleven toes and various optical trickery conspires to make it look like I have ten.

If this seems worrying to you, consider that you may be asking too much of this pre-rationality condition. After all, if you have a weird prior, you have a weird prior. In the machine-generating-random-priors world, you already believe that your prior is a good fit for the world. That's what it is to have a prior. Yes, according to our *actual* posteriors it seems like there should be no correlation between these random priors and the world they're in, but asking the pre-rationality condition to make our actual beliefs win out seems like a pretty illicit move.

Another worry is that it seems there's some spooky action-at-a-distance going on between the pre-rationality condition and the assignment of priors. Once everyone has their priors, the pre-rationality condition is powerless to change them. So how is the pre-rationality condition making it so that everyone has the same prior?

I claim that actually, this presentation of the pre-Bayesian proof is not quite right. According to me, if I'm a *Bayesian* and believe our priors are equally good, then we must have the same priors. If I'm a pre-Bayesian and believe our priors are equally good, then I must *believe* that your prior averages out to mine. This latter move is open to the pre-Bayesian (who has uncertainty about priors) but not to the Bayesian (who knows the priors).

I'll make an argument purely within Bayesianism for believing in equally good priors to having the same prior, and then we'll see how belief in priors comes in for a pre-Bayesian.

Bayesian prior equality

To get this off the ground, I want to make precise the claim of believing someone's priors are as good as yours. I'm going to look at 3 ways of doing this. Note that Hanson doesn't suggest a particular one, so he doesn't have to accept any of these as what he means, and that might change how well my argument works.

Let's suppose my prior is p and yours is q . Note, these are fixed functions, not references pointing at my prior and your prior. In the Bayesian framework, we just have our priors, end of story. We don't reason about cases where our priors were different.

Let's suppose *score* is a strictly proper scoring rule (if you don't know what that means, I'll explain in a moment). *score* takes in a probability distribution over a random variable and an actual value for that random variable. It gives more points the more of the probability distribution's mass is near the actual value. For it to be strictly proper, I uniquely maximise my expected score by reporting my true probability distribution. That is $E_p[\text{score}(f, X)]$ is uniquely maximised when $f = p$.

Let's also suppose my posterior is $p|B$, that is (using notation a bit loosely) my prior probability conditioned on some background information B .

Here are some attempts to precisely claim someone's prior is as good as mine:

1. For all X , $E_p[\text{score}(p, X)] = E_p[\text{score}(q, X)]$.
2. For all X , $E_{p|B}[\text{score}(p|B, X)] = E_{p|B}[\text{score}(q|B, X)]$.
3. For all X , $E_{p|B}[\text{score}(p, X)] = E_{p|B}[\text{score}(q, X)]$.

(1) says that, according to my prior, your prior is as good as mine. By the definition of a proper scoring rule, this means that your prior is the same as mine.

(2) says that, according to my posterior, the posterior you'd have with my current information is as good as the posterior I have. By the definition of the proper scoring rule, this means that your posterior is equal to my posterior. This is a bit broader than (1), and allows your prior to have already "priced in" some information that I now have.

(3) says that given what we know now, your prior was as good as mine.

That rules out $q = p|B$. That would be a prior that's better than mine: it's just what you get from mine when you're already certain you'll observe some evidence (like an apple falling in 1663). Observing that evidence doesn't change your beliefs.

In general, it can't be the case that you predicted B as more likely than me, which can be seen by taking $X = B$.

On future events, your prior can match my prior, or diverge from my posterior equally as far as my prior, but in the opposite direction.

I don't really like 3, because while it accepts that your prior was as good as mine in the past, it can think that after you update your prior you'll still be worse than me.

That leaves us with 1 and 2 then. If 1 or 2 are our precise notion, then it follows quickly that we have common priors.

This is just a notion of logical consistency though; I don't have room for believing that our prior-generating processes make yours as likely to be true as mine. It's just that if the probability distribution that happens to be your prior appears to me as good as the probability distribution that happens to be my prior, they are the same probability distribution.

Pre-Bayesian prior equality

How to make pre-Bayesian claim that your prior is as good as mine?

Here let, p_i be my prior as a reference, rather than as a concrete probability distribution. Claims about p_i are claims about my prior, no matter what function that actually ends up being. So for example, claiming that p_i scores well is claiming that as we look at different worlds, we see it is likely that my prior is a well-adapted prior for that specific world. In contrast, a claim that p scores well would be a claim that the actual world looks a lot like p .

Similarly, p_j is your prior as a reference. Let \mathbf{p} be a vector assigning a prior to each agent.

Let f be my pre-prior. That is, my initial beliefs over combinations of worlds and prior assignments. Similarly to above, let $f|B$ be my pre-posterior (a bit of an awkward term, I admit).

For ease of exposition (and I don't think entirely unreasonably), I'm going to imagine that I know my prior precisely. That is $f(w, p) = 0$ if $p_i \neq p$.

Here are some ways of making the belief that your prior is as good as mine precise in the pre-Bayesian framework.

1. For all X , $E_p[\text{score}(p, X)] = E_f[\text{score}(p, X)]$.
2. For all X , $E_{p|B}[\text{score}(p|B, X)] = E_{f|B}[\text{score}(p|B, X)]$.
3. For all X , $E_{p|B}[\text{score}(p, X)] = E_{f|B}[\text{score}(p, X)]$.

On the LHS, the expectation uses p rather than f , because of the pre-rationality condition. Knowing my prior, my updated pre-prior agrees with it about the probability of the ground events. But I still don't know your prior, so I have to use f on the RHS to "expect" over the event and your prior itself.

(1) says that, according to my pre-prior, your prior is as good as mine in expectation. The proper scoring rule says that my prior is the unique maximum *for a fixed function*. But I could, in principle, believe that your prior is better adapted to each world than my prior, but I'm still not certain which world we're in (or what your prior is), so I can't update my beliefs.

Given the equality, I can't want to switch priors with you in general, but I could think you have a prior that's more correlated with truth than mine in some cases and less so in others.

(2) says that, according to my pre-posterior, your prior conditioned on my info is, in expectation, as good as my prior conditioned on my info.

I like this better than (1). Evidence in the real world leads me to beliefs about the prior production mechanisms (like genes, nurture and so on). These don't seem to give a good reason for my innate beliefs to be better than anyone else's. Therefore, I believe your prior is probably as good as mine on average.

But note, I don't actually know what your prior is. It's just that *I believe* we probably share similar priors. The spooky action-at-a-distance is eliminated. This is just (again) a claim about consistent beliefs: if I believe that your prior got generated in a way that made it as good as mine, then I must believe it's not too divergent from mine.

3. says that, given what we now know, I think your prior is no better or worse than mine in expectation. This is about as unpalatable in the pre-Bayesian as the Bayesian case.

So, on either (1) or (2), I believe that your prior will, on average, do as well as mine. I may not be sure what your prior is, but cases where it's far better will be matched by cases where it's far worse. Even knowing that your prior performs exactly as well as mine, I might not know exactly which prior you have. I know that all the places it does worse will be matched by an equal weight of places where it does better, so I can't appeal to my prior as a good reason for us to diverge.

When does rationality-as-search have nontrivial implications?

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

(This originated as [a comment](#) on the post "[Embedded World-Models,](#)" but it makes a broadly applicable point and is substantial enough to be a post, so I thought I'd make it a post as well.)

[This post](#) feels quite similar to things I have written in the past to justify my lack of enthusiasm about idealizations like AIXI and logically-omniscient Bayes. But I would go further: I think that grappling with embeddedness properly will *inevitably* make theories of this *general type* irrelevant or useless, so that "a theory like this, except for embedded agents" is not a thing that we can reasonably want. To specify what I mean, I'll use this paragraph as a jumping-off point:

Embedded agents don't have the luxury of stepping outside of the universe to think about how to think. What we would like would be a theory of rational belief for *situated* agents which provides foundations that are similarly as strong as the foundations Bayesianism provides for dualistic agents.

Most "theories of rational belief" I have encountered -- including Bayesianism in the sense I think is meant here -- are framed at the level of an evaluator outside the universe, and have essentially no content when we try to transfer them to individual embedded agents. This is because these theories tend to be derived in the following way:

- We want a theory of the best possible behavior for agents.
- We have some class of "practically achievable" strategies S , which can actually be implemented by agents. We note that an agent's observations provide some information about the quality of different strategies $s \in S$. So if it were possible to follow a rule like $R \equiv$ "find the best $s \in S$ given your observations, and then follow that s ," this rule would spit out very good agent behavior.
- Usually we soften this to a performance-weighted average rather than a hard argmax, but the principle is the same: if we could search over all of S , the rule R that says "do the search and then follow what it says" can be competitive with the very best $s \in S$. (Trivially so, since it has access to the best strategies, along with all the others.)
- But usually $R \notin S$. That is, the strategy "search over all practical strategies and follow the best ones" is not a *practical* strategy. But we argue that this is fine, since we are constructing a theory of *ideal* behavior. It doesn't have to be practically implementable.

For example, in Solomonoff, S is defined by computability while R is allowed to be uncomputable. In the LIA construction, S is defined by polytime complexity while R is allowed to run slower than polytime. In logically-omniscient Bayes, finite sets of hypotheses can be manipulated in a finite universe but the full Boolean algebra over hypotheses generally cannot (N.B. I don't think this last case fits my schema quite as well as the other two).

I hope the framework I've just introduced helps clarify what I find unpromising about these theories. By construction, any agent you can actually design and run is a *single* element of S (a "practical strategy"), so every fact about rationality that can be incorporated into agent design gets "hidden inside" the individual $s \in S$, and the only things you can learn from the "ideal theory" R are things which can't fit into a practical strategy.

For example, suppose (reasonably) that model averaging and complexity penalties are broadly good ideas that lead to good results. But all of the model averaging and complexity penalization that can be done *computably* happens inside some Turing machine or other, at the level "below" Solomonoff. Thus Solomonoff *only* tells you about the extra advantage you can get by doing these things *uncomputably*. Any kind of nice Bayesian average over Turing machines that can happen computably is (of course) just another Turing machine.

This also explains why I find it misleading to say that good practical strategies constitute "approximations to" an ideal theory of this type. Of course, since R just says to follow the best strategies in S , if you are following a very good strategy in S your behavior will tend to be close to that of R . But this cannot be attributed to *any* of the searching over S that R does, since you are not doing a search over S ; you are executing a *single member* of S and ignoring the others. Any searching that can be done practically collapses down to a single practical strategy, and any that doesn't is not practical.

Concretely, this talk of approximations is like saying that a very successful chess player "approximates" the rule "consult all possible chess players, then weight their moves by past performance." Yes, the skilled player will *play similarly* to this rule, but they are not *following* it, not even approximately! They are only themselves, not any other player.

Any theory of ideal rationality that wants to be a guide for embedded agents will have to be constrained in the same ways the agents are. But theories of ideal rationality usually get *all of their content* by going to a level above the agents they judge. So this new theory would have to be a very different sort of thing.

To state all this more pithily: if we design the search space to contain everything feasible, then rationality-as-search has no feasible implications. If rationality-as-search

is to have feasible implications, then the search space must be weak enough for there to be *something* feasible that is *not* a point in the search space.

Aligned AI, The Scientist

The problem with constructing an aligned AI is that any active utility function or attempt at world optimization is likely to succumb to the Goodhart's law in one of its many forms, as discussed here and elsewhere by the good people of MIRI. I wonder if a more passive approach is worth considering, or may have been considered already.

Humanity is a part of the Universe, and building and organizing accurate knowledge about the Universe is what science is. Not using the scientific knowledge to advance specific interests or goals, e.g. technological advancements or personal gain, but for the knowledge's sake. Such a scientifically-minded agent would not be interested in modifying the Universe, and would limit any effects to the minimum needed to understand it. A part of this scientific research would be to understand humanity as deeply as possible, including what we humans imagine an aligned AI would look like even though we do not fully understand it ourselves at this point.

Presumably at some point such an AI would understand the universe and the humans in it enough to basically serve as a safe DWIM (do what I mean) genie. It would be inherently safe because doing anything unsafe, or agreeing to do anything unsafe would mean that the genie does not understand the part of the Universe that is the humanity. After all, we would not want to do anything that has unsafe and unintended consequences. "Unsafe" includes doing nothing at all: an AI that would prevent humans from doing anything would not understand humans, and so would not understand the universe. In other words

Aligned AI is AI the scientist, not AI the engineer.

This is, of course, is easier said than done. Learning all about the world while actively minimizing any impact on the world is something that we humans often strive to do when trying to understand the ecosystem of the Earth, with mixed results. Still, sometimes we succeed, and, odds are, so could an agent smarter than us.

Fixed Point Exercises

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

Sometimes people ask me what math they should study in order to get into agent foundations. My first answer is that I have found the introductory class in every subfield to be helpful, but I have found the later classes to be much less helpful. My second answer is to learn enough math to understand [all fixed point theorems](#). These two answers are actually very similar. Fixed point theorems span all across mathematics, and are central to (my way of) thinking about agent foundations.

This post is the start of a sequence on fixed point theorems. It will be followed by several posts of exercises that use and prove such theorems. While these exercises aren't directly connected to AI safety, I think they're quite useful for preparing to think about agent foundations research. Afterwards, I will discuss the core ideas in the theorems and where they've shown up in alignment research.

The math involved is not much deeper than a first course in the various subjects (logic, set theory, topology, computability theory, etc). If you don't know the terms, a bit of googling, wikipedia and math.stackexchange should easily get you most of the way. Note that the posts can be tackled in any order.

Here are some ways you can use these exercises:

- You can host a local MIRIx group, and go through the exercises together. This might be useful to give a local group an affordance to work on math rather than only reading papers.
- You can work on them by yourself for a while, and post questions when you get stuck. You can also post your solutions to help others, let others see an alternate way of doing a problem, or help you realize that there is a problem with your solution.
- You can skip to the discussion (which has some spoilers), learn a bunch of theorems from Wikipedia, and use this as a starting point for trying to understand some MIRI papers.
- You can use answering these questions as a goalpost for learning a bunch of introductory math from a large collection of different subfields.
- You can show off by pointing out that some of the questions are wrong, and then I will probably fix them and thank you.

The first set of exercises is [here](#).

Thanks to Sam Eisenstat for helping develop these exercises, Ben Pace for helping edit the sequence, and many AISFP participants for testing them and noticing errors.

Meta

Read the following.

Please use the (new) spoilers feature - the symbol '>' followed by '!' followed by space - in your comments to hide all solutions, partial solutions, and other discussions of the math. The comments will be moderated strictly to cover up spoilers!

I recommend putting all the object level points in spoilers and leaving metadata outside of the spoilers, like so:

Here's my solution / partial solution / confusion for question #5:

And put your idea in here! (reminder: LaTex is cmd-4 / ctrl-4)

Four factors that moderate the intensity of emotions

Epistemic status: influenced by my study of emotions over the years, but primarily based on personal observation. Presented in the spirit of "here's a model, judge for yourself whether you think it's any good."

If you're short on time or just skimming, I suggest skipping to the description of Closeness-of-the-counterfactual. It's the factor I think is least recognized and the one I most wanted to write about.

What makes some emotions stronger than others? Why are some the faintest whispers, easily missed and others roaring, crashing storms which threaten to consume us?

The obvious answer is that emotions vary in intensity in proportion to the *magnitude* of what they're about. Things which are a little bit good or a little bit bad evoke weak pleasant or aversive feelings, while things which are amazingly good or terribly bad provoke strong feeling. However, I assert that this *magnitude* is only one factor among several and is insufficient on its own to explain what causes strong or weak emotions.

In this post, I list the factors which are salient to me: magnitude plus three others. I do not think that they are especially surprising or profound, but I claim paying attention to them allows us to better appreciate the mechanistic and lawful operation of emotions. This appreciation is practical in that it lets us better recognize and remedy common emotional pathologies. [See *Section 2: Problematic manipulations of the factors*].

Section 1: The Factors

- **Magnitude [of the stimulus]**
- **Attention**
- **Closeness-of-the-counterfactual**
- **Actionability**

Magnitude [of the stimulus]

This factor is the most obvious and least profound of all of the factors. However, it is illuminating to note just how insufficient it is to drive an emotion in the absence of the other factors.

Taking it as an assumption that all emotions are about something in the world [1], the strength of the emotion generally scales with *magnitude* of the “goodness” or “badness” which caused it. One feels stronger grief when their house burns down than when they dropped their cookie in the dirt. Making thousands from Bitcoin feels better than finding a twenty dollar bill on the street.

Attention

Obvious and yet still underappreciated.

Any given person is aware of thousands of situations, circumstances, and facts that could evoke just about any feeling. Meditate on your good fortune and you might feel happiness; think about those starving and diseased and you’ll feel sad; remember that unfair thing your teacher did in third grade and you’ll feel mad, and so on.

Emotions are usually *about* reality, but the emotions we experience are not about all the realities of which we are knowledgeable. No human could ever contain that much emotion. Instead, our emotions tend to be about *whatever happens to be in our attention (broadly defined)*.

It’s like humans have this “attention slot” where you can put something, i.e. think about it, and then that’s what you’ll have emotions about. I am using attention in a broad and loose sense here. What I’m gesturing isn’t fully under one’s control and extends beyond conscious awareness. Think of how strong grief can stay present somewhere in your mind even while you try to do other things.

That we have emotions only about things in the “attention slot” solves the problem of humans not being able to simultaneously have emotions about all the realities of which they are aware (or could imagine), but more importantly it serves the adaptive purpose of emotions. **Emotions are meant to guide behavior. It makes sense that emotions should be driven by the immediate, contextual, things we’re currently dealing with, i.e. those things we’re paying attention to.** It’s not valuable to be feeling happy about something good which happened a month ago if

right now you're in a bad situation you should get out of. Your emotions will be about the current situation, assuming that is, you're paying attention to your present.

Even if you're ruminating about your past, it should be [2] so that you can learn from mistakes and successes in a healthy way so as to succeed going forward. If you are fantasizing about possible futures, it should be to motivate you to work towards them.

The attention-moderator nature of emotion gives rise to a number of common observations:

- We're less upset by things over time; they're no longer events we're paying attention to and they're less relevant.
- People attempt to feel better by distracting themselves from upsetting circumstances [see below *Section 2: Problematic manipulation of these factors*].
- [Gratitude journaling](#) makes people happier.
- We can evoke emotions just by thinking about things, i.e. placing our attention on them, regardless of whether they're real or imaginary, good or bad.

Closeness-of-the-counterfactual

Despite its supreme importance, this is the most under-recognized moderator of emotion intensity. As far as I'm aware, not that I've scoured the psych literature, there isn't a common name for it.

ETA: [Gram Stone](#) points out that [Kahneman & Tversky \(1981\)](#) describe this concept despite not coining a term for it, and that [Roese \(1997\)](#) provides a good early review on counterfactual thinking including contrasting effects similar to what is discussed here.

For the most part, it is a lot more frustrating to have missed your flight by five minutes than it is to miss your flight by five hours. It is a lot more frustrating to miss your flight when it seems you could have changed one or two small actions to have made it rather than when success was just out of your control. For example, it is more frustrating to miss the flight if the cause was you spent too long on Facebook rather than your car happened to be improbably stolen right from your garage.

It is more disappointing to not get a job when you thought you would, and more exhilarating to win a competition if you weren't sure you'd win.

In general, emotions which relate to a [counterfactual](#) (i.e., how things might have been different, i.e., pretty much all emotions), scale in intensity in proportion to how easy it is to imagine [3] the counterfactual having been true. It's easier to imagine having made your flight when it would have taken a small decision on your part to make the difference, and somewhat harder if something out of your control would have needed happen differently or it would have taken an improbably large amount of effort on your part.

I call the "how easy it is to imagine the counterfactual" property *closeness-of-the-counterfactual*, or *counterfactual-closeness* for short.

As usual, counterfactual-closeness as a moderator of emotion intensity makes sense if emotions are supposed to be adaptive. It's adaptive to have strong emotions about

counterfactual realities you could have nearly reached - if only you'd done a few things different - than about realities, no matter how pleasant, that never seemed in reach. It just doesn't get me anywhere to be dreadfully sad all the time that I wasn't born able to fly.

It's a simple principle, yet is consistent with many observations beyond the above.

- We are more upset by things that people around us have than things no one has. I am sad that I don't have a swimming pool when my neighbors do and I could technically afford it, but not sad that I don't have a spaceship because that just seems unrealistic - unless I'm Musk or Bezos, I have no reasonable expectation that I would have one.
- We are more sad to lose things we once had than about things we've never had.
- Unexpected good fortune feels a lot more rewarding than expected good fortune.
- Sometimes people try to feel better about a failure by saying "it was always hopeless." In effect, they are trying to create counterfactual-distance so there is less pain.
 - This is related to a "sour grapes" response.
- It is consistent with [Eliezer's observation](#) that most people can't find motivation to do things they think are less than 70% likely to succeed. Perhaps you need to assign 70% probability of success to have enough counterfactual-closeness to evoke emotion.
- Vivid descriptions of things, images, and videos cause us to have stronger emotions. The representations make them easier to imagine (and also help load them into attention).

Probably related: [Book Review: Surfing Uncertainty](#)

Actionability

Continuing the theme that emotions ought to be adaptive, it makes a lot more sense to have emotions about situations where you can do something than ones where you can't, or at least about situations where you could have done something different.

Even in cases where an emotion [might seem inert](#), the emotion itself is probably trying to effect the world.

However, I'm not as sure of my understanding of this moderator as the others. It might actually bring about more of a qualitative difference in emotions than quantitative. The observation I'm drawing on here is that the emotions related to unresolved conflict with a colleague feel different from those attending grief about something which is done and dusted. In the former, there's a "pulling" from the emotion as though it wants something, while in the latter the emotional tone is clear and pure, just signalling to my mind that something bad happened and I should do things different in the future.

Section 2: Problematic manipulations of the factors

Humans are crafty creatures with awareness of their own minds and the ability to manipulate the inputs they feed into their own minds to game the system. **In short, we have some ability to wirehead.** Or even if we're not wireheading, these factors are pieces of the system which can be vulnerable to specific attacks or their own failure modes.

Attention and counterfactual-closeness are clear examples.

Manipulating attention: distraction to avoid unpleasant emotions

It's easy to see that many people distract themselves from unpleasant realities to escape unpleasant emotions. I think it's underappreciated just how absurdly widespread the behavior is.

The pernicious part of avoidance-behaviors is that many behaviors used for distraction, i.e. almost anything pleasurable, could be done purely for the sake of pleasure and in many cases is perfectly healthy and good. It's easy to claim that you're eating cake simply because cake is tasty unrelated to anything else. Yet, people are often compulsively and habitually looking for something stimulating to keep attention off the painful [4].

Compulsion is likely the differentiator about whether pleasure seeking behavior is driven by distraction and avoidance. Is a person reading a novel because they really feel like it and it's a good time, or is there something it is helping them avoid, e.g. an assignment? The test I apply in this case is to ask whether I think a particular behavior optimizes my life as a whole or whether it's just this experience in the moment being optimized. Over what timescale does this behavior improve on my life? Manipulation of attention to avoid pain (wireheading) will often be to the detriment of one's life overall - pleasant in the short run, worse in the long-run.

Manipulating counterfactual-closeness

There is a temptation to artificially increase counterfactual-closeness to realities which are pleasant. Someone might cling to a dream of becoming an olympic runner, even as evidence mounts against them. They focus solely on the positive signs and they repeatedly and obsessively enumerate the pathways through which it all might work out. They've distorted their view of the world to see the desired world as much closer to reality than it is, because it feels good. They even become attached to the fantasy they've constructed. To maintain, it they have to twist the evidence and twist their epistemics, i.e. a one-sided counting of all evidence in one direction while ignoring all

contrary data. This behavior is common in romantic contexts too. I assert that overall people behaving this way would be better served by good epistemics and accurate assessments of counterfactual-closeness.

There is equally a motivation to decrease counterfactual-closeness, i.e., increase counterfactual-distance. Things are easier to accept when they seem necessary, unavoidable, and so believing them so is a way to avoid pain. Most people do not appear to be pained that millions of people are dying, millions starving, millions diseased. They're not distressed by their own imminent and assured death. Partly I think the pain is avoided by avoiding place attention on these topics, but also I think there's motivated cognition to believe that it is impossible to do anything. Merely believing that there is something which could be done, having greater counterfactual-closeness, means experiencing pain that the something hasn't been done yet.

This explains that people are resistant if you try to tell them things they could do. Believing something could be done would require them to move counter to a hedonic gradient, out of a local optimum. Believing something could be done (but hasn't been) hurts more than believing nothing could be done, even though the former is how you get the best state of all - where something has been done successfully.

Manipulating magnitude and actionability

I imagine that people will readily recognize the behavior of people protesting through tears that something is "no big deal" in attempt to minimize their feelings - they are minimizing magnitude. And the behavior of insisting "nothing can be done" to quieten any nagging sense of responsibility, external or internal - they are minimizing actionability.

Endnotes

[1] Arguably some might emotions are not about anything, e.g. in emotion dysregulation disorders such as depression. My counter is that even if some emotions are detached from reality, that is a breakdown in the proper operation whose design and purpose is guide behavior within reality. Any "healthy" emotion will be "about" something.

[2] "Should" from the perspective of what emotions are "designed for", namely that they are trying to drive adaptive action..

[3] The relevant kind of "imagination" here is a *S1*, *instinctive* feeling around gut expectations about what will happen or could have happened. It's more than a *S2*, abstract picturing of a scenario in your mind.

[4] When I say "painful", I mean anything at all slightly aversive. If I'm shy and dislike phone calls, I might put off calling the bank about the mistaken charge for weeks to avoid my slight discomfort. Us humans are sensitive to even the gentlest hedonic gradients.

October gwern.net links

This is a linkpost for <https://www.gwern.net/newsletter/2018/10>

Diagonalization Fixed Point Exercises

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

This is the second of three sets of fixed point exercises. The first post in this sequence is [here](#), giving context.

1. Recall Cantor's diagonal argument for the uncountability of the real numbers.
Apply the same technique to convince yourself that for any set S , the cardinality of S is less than the cardinality of the power set $P(S)$ (i.e. there is no surjection from S to $P(S)$).
2. Suppose that a nonempty set T has a function f from T to T which lacks fixed points (i.e. $f(x) \neq x$ for all $x \in T$). Convince yourself that there is no surjection from S to $S \rightarrow T$, for any nonempty S . (We will write the set of functions from A to B either as $A \rightarrow B$ or B^A ; these are the same.)
3. For nonempty S and T , suppose you are given $g : S \rightarrow T^S$ a surjective function from the set S to the set of functions from S to T , and let f be a function from T to itself. The previous result implies that there exists an x in T such that $f(x) = x$.
Can you use your proof to describe x in terms of f and g ?
4. Given sets A and B , let $\text{Comp}(A, B)$ denote the space of total computable functions from A to B . We say that a function from C to $\text{Comp}(A, B)$ is computable if and only if the corresponding function $f' : C \times A \rightarrow B$ (given by $f'(c, a) = f(c)(a)$) is computable. Show that there is no surjective computable function from the set S of all strings to $\text{Comp}(S, \{T, F\})$.
5. Show that the previous result implies that there is no computable function $\text{halt}(x, y)$ from $S \times S \rightarrow \{T, F\}$ which outputs T if and only if the first input is a code for a Turing machine that halts when given the second input.

6. Given topological spaces A and B, let $\text{Cont}(A, B)$ be the space with the set of continuous functions from A to B as its underlying set, and with topology such that $f : C \rightarrow \text{Cont}(A, B)$ is continuous if and only if the corresponding function $f' : C \times A \rightarrow B$ (given by $f'(c, a) = f(c)(a)$) is continuous, assuming such a space exists. Convince yourself that there is no space X which continuously surjects onto $\text{Cont}(X, S)$, where S is the circle.
7. In your preferred programming language, write a quine, that is, a program whose output is a string equal to its own source code.
8. Write a program that defines a function f taking a string as input, and produces its output by applying f to its source code. For example, if f reverses the given string, then the program should outputs its source code backwards.
9. Given two sets A and B of sentences, let $\text{Syn}(A, B)$ be the set of all functions from A to B defined by substituting the Gödel number of a sentence in A into a fixed formula. Let S_0 be the set of all sentences in the language of arithmetic with one unbounded universal quantifier and arbitrarily many bounded quantifiers, and let S_1 be the set of all formulas with one free variables of that same quantifier complexity. By representing syntax using arithmetic, it is possible to give a function $f \in \text{Syn}(S_1 \times S_1, S_0)$ that substitutes its second argument into its first argument. Pick some coding of formulas as natural numbers, where we denote the number coding for a formula ϕ as $\Gamma\psi\Gamma$. Using this, show that for any formula $\phi \in S_1$, there is a formula $\psi \in S_0$ such that $\phi(\Gamma\psi\Gamma) \leftrightarrow \psi$.
10. (Gödel's second incompleteness theorem) In the set S_1 , there is a formula $\neg\text{Bew}$ such that $\neg\text{Bew}(\Gamma\psi\Gamma)$ holds iff the sentence ψ is not provable in Peano arithmetic. Using this, show that Peano arithmetic cannot prove its own consistency.
11. (Löb's theorem) More generally, the diagonal lemma states that for any formula ϕ with a single free variable, there is a formula ψ such that, provably, $\phi(\Gamma\psi\Gamma) \leftrightarrow \psi$. Now, suppose that Peano arithmetic proves that $\text{Bew}(\psi) \rightarrow \psi$ for some formula ψ . Show that Peano arithmetic also proves ψ itself. Some facts that you may need

are that (a) when a sentence ψ is provable, the sentence $\text{Bew}(\psi)$ is itself provable, (b) Peano arithmetic proves this fact, that is, Peano arithmetic proves $\text{Bew}(\psi) \rightarrow \text{Bew}(\text{Bew}(\psi))$, for any sentence ψ and (c) Peano arithmetic proves the fact that if χ and $\chi \rightarrow \zeta$ are provable, then ζ is provable.

12. (Tarski's theorem) Show that there does not exist a formula ϕ with one free variable such that for each sentence ψ , the statement $\phi(\Gamma\psi\Gamma) \leftrightarrow \psi$ holds.

13. Looking back at all these exercises, think about the relationship between them.

Please use the spoilers feature - the symbol '>' followed by '!' followed by space -in your comments to hide all solutions, partial solutions, and other discussions of the math. The comments will be moderated strictly to hide spoilers!

I recommend putting all the object level points in spoilers and including metadata outside of the spoilers, like so: "I think I've solved problem #5, here's my solution <spoilers>" or "I'd like help with problem #3, here's what I understand <spoilers>" so that people can choose what to read.

Alignment Newsletter #33

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

Find all Alignment Newsletter resources [here](#). In particular, you can [sign up](#), or look through the [database](#) of all summaries.

One correction to last week's newsletter: the title *Is Robustness at the Cost of Accuracy* should have been *Is Robustness the Cost of Accuracy*.

Highlights

[**Reward learning from human preferences and demonstrations in Atari**](#) (*Borja Ibarz et al*): We have had lots of work on learning from preferences, demonstrations, proxy rewards, natural language, rankings etc. However, most such work focuses on one of these modes of learning, sometimes combined with an explicit reward function. This work learns to play Atari games using both preference and demonstration information. They start out with a set of expert demonstrations which are used to initialize a policy using behavioral cloning. They also use the demonstrations to train a reward model using the DQfD algorithm. They then continue training the reward and policy simultaneously, where the policy is trained on rewards from the reward model, while the reward model is trained using preference information (collected and used in the same way as Deep RL from Human Preferences) and the expert demonstrations. They then present a *lot* of experimental results. The main thing I got out of the experiments is that when demonstrations are good (near optimal), they convey a lot of information about how to perform the task, leading to high reward, but when they are not good, they will actively hurt performance, since the algorithm assumes that the demonstrations are high quality and the demonstrations "override" the more accurate information collected via preferences. They also show results on efficiency, the quality of the reward model, and the reward hacking that can occur if you don't continue training the reward model alongside the policy.

Rohin's opinion: I'm excited to see work that combines information from multiple sources! In general with multiple sources you have the problem of figuring out what to do when the sources of information conflict, and this is no exception. Their approach tends to prioritize demonstrations over preferences when the two conflict, and so in cases where the preferences are better (as in Enduro) their approach performs poorly. I'm somewhat surprised that they prioritize demos over preferences, since it seems humans would be more reliable at providing preferences than demos, but perhaps they needed to give demos more influence over the policy in order to have the policy learn reasonably quickly. I'd be interested in seeing work that tries to use the demos as much as possible, but detect when conflicts happen and prioritize the preferences in that situation -- my guess is that this would let you get good performance across most Atari games.

Technical AI alignment

Embedded agency sequence

[Embedded Agency \(full-text version\)](#) (*Scott Garrabrant and Abram Demski*): This is the text version of all of the previous posts in the sequence.

Iterated amplification sequence

[The Steering Problem](#) (*Paul Christiano*): The steering problem refers to the problem of writing a program that uses black-box human-level cognitive abilities to be as useful as a well-motivated human Hugh (that is, a human who is "trying" to be helpful). This is a conceptual problem -- we don't have black-box access to human-level cognitive abilities yet. However, we can build suitable formalizations and solve the steering problem within those formalizations, from which we can learn generalizable insights that we can apply to the problem we will actually face once we have strong AI capabilities. For example, we could formalize "human-level cognitive abilities" as Hugh-level performance on question-answering (yes-no questions in natural language), online learning (given a sequence of labeled data points, predict the label of the next data point), or embodied reinforcement learning. A program P is more useful than Hugh for X if, for every project using a simulation of Hugh to accomplish X, we can efficiently transform it into a new project which uses P to accomplish X.

Rohin's opinion: This is an interesting perspective on the AI safety problem. I really like the ethos of this post, where there isn't a huge opposition between AI capabilities and AI safety, but instead we are simply trying to figure out how to use the (helpful!) capabilities developed by AI researchers to do useful things.

If I think about this from the perspective of reducing existential risk, it seems like you also need to make the argument that AI systems are unlikely to pose an existential threat before they are human-level (a claim I mostly agree with), or that the solutions will generalize to sub-human-level AI systems.

[Clarifying "AI Alignment"](#) (*Paul Christiano*): I previously summarized this in [AN #2](#), but I'll consider it in more detail now. As Paul uses the term, "AI alignment" refers only to the problem of figuring out how to build an AI that is *trying* to do what humans want. In particular, an AI can be aligned but still make mistakes because of incompetence. This is not a formal definition, since we don't have a good way of talking about the "motivation" of an AI system, or about "what humans want", but Paul expects that it will correspond to some precise notion after we make more progress.

Rohin's opinion: Ultimately, our goal is to build AI systems that reliably do what we want them to do. One way of decomposing this is first to *define* the behavior that we want from an AI system, and then to figure out how to obtain that behavior, which we might call the definition-optimization decomposition. [Ambitious value learning](#) aims to solve the definition subproblem. I interpret this post as proposing a *different decomposition* of the overall problem. One subproblem is how to build an AI system that is *trying* to do what we want, and the second subproblem is how to make the AI competent enough that it *actually* does what we want. I like this motivation-competence decomposition for a few reasons, which I've written a [long comment](#) about that I strongly encourage you to read. The summary of that comment is: motivation-competence isolates the urgent part in a single subproblem (motivation), humans are an existence proof that the motivation subproblem can be solved, it is possible to apply the motivation framework to systems without lower capabilities, the safety guarantees degrade slowly and smoothly, the definition-optimization decomposition as exemplified by expected utility maximizers has generated primarily negative results, and motivation-competence allows for

interaction between the AI system and humans. The major con is that the motivation-competence decomposition is informal, imprecise, and may be intractable to work on.

[An unaligned benchmark](#) (*Paul Christiano*): I previously summarized this in [Recon #5](#), but I'll consider it in more detail now. The post argues that we could get a very powerful AI system using model-based RL with MCTS. Specifically, we learn a generative model of dynamics (sample a sequence of observations given actions), a reward model, and a policy. The policy is trained using MCTS, which uses the dynamics model and reward model to create and score rollouts. The dynamics model is trained using the actual observations and actions from the environment. The reward is trained using preferences or rankings (think something like [Deep RL from Human Preferences](#)). This is a system we could program now, and with sufficiently powerful neural nets, it could outperform humans.

However, this system would not be aligned. There could be specification failures: the AI system would be optimizing for making humans think that good outcomes are happening, which may or may not happen by actually having good outcomes. (There are a few arguments suggesting that this is likely to happen.) There could also be robustness failures: as the AI exerts more control over the environment, there is a distributional shift. This may lead to the MCTS finding previously unexplored states where the reward model accidentally assigns high reward, even though it would be a bad outcome, causing a failure. This may push the environment even more out of distribution, triggering other AI systems to fail as well.

Paul uses this and other potential AI algorithms as *benchmarks* to beat -- we need to build aligned AI algorithms that achieve similar results as these benchmarks. The further we are from hitting the same metrics, the larger the incentive to use the unaligned AI algorithm.

Iterated amplification could potentially solve the issues with this algorithm. The key idea is to always be able to cash out the learned dynamics and reward models as the result of (a large number of) human decisions. In addition, the models need to be made robust to worst case inputs, possibly by using [these techniques](#). In order to make this work, we need to make progress on robustness, amplification, and an understanding of what bad behavior is (so that we can argue that it is easy to avoid, and iterated amplification does avoid it).

Rohin's opinion: I often think that the hard part of AI alignment is actually the *strategic* side of it -- even if we figure out how to build an aligned AI system, it doesn't help us unless the actors who actually build powerful AI systems use our proposal. From that perspective, it's very important for any aligned systems we build to be competitive with unaligned ones, and so keeping these sorts of benchmarks in mind seems like a really good idea. This particular benchmark seems good -- it's essentially the AlphaGo algorithm, except with learned dynamics (since we don't know the dynamics of the real world) and rewards (since we want to be able to specify arbitrary tasks), which seems like a good contender for "powerful AI system".

Fixed point sequence

[Fixed Point Exercises](#) (*Scott Garrabrant*): Scott's advice to people who want to learn math in order to work on agent foundations is to learn all of the fixed-point theorems across the different areas of math. This sequence will present a series of exercises designed to teach fixed-point theorems, and will then talk about core ideas in the theorems and how the theorems relate to alignment research.

Rohin's opinion: I'm not an expert on agent foundations, so I don't have an opinion worth saying here. I'm not going to cover the posts with exercises in the newsletter -- visit the [Alignment Forum](#) for that. I probably will cover the posts about how the theorems relate to agent foundations research.

Agent foundations

[Dimensional regret without resets](#) (*Vadim Kosoy*)

Learning human intent

[Reward learning from human preferences and demonstrations in Atari](#) (*Borja Ibarz et al*): Summarized in the highlights!

[Acknowledging Human Preference Types to Support Value Learning](#) (*Nandi, Sabrina, and Erin*): Humans often have multiple "types" of preferences, which any value learning algorithm will need to handle. This post concentrates on one particular framework -- liking, wanting and approving. Liking corresponds to the experience of pleasure, wanting corresponds to the motivation that causes you to take action, and approving corresponds to your conscious evaluation of how good the particular action is. These correspond to different data sources, such as facial expressions, demonstrations, and rankings respectively. Now suppose we extract three different reward functions and need to use them to choose actions -- how should we aggregate the reward functions? They choose some desiderata on the aggregation mechanism, inspired by social choice theory, and develop a few aggregation rules that meet some of the desiderata.

Rohin's opinion: I'm excited to see work on dealing with conflicting preference information, particularly from multiple data sources. To my knowledge, there isn't any work on this -- while there is work on multimodal input, usually those inputs don't conflict, whereas this post explicitly has several examples of conflicting preferences, which seems like an important problem to solve. However, I would aim for a solution that is less fixed (i.e. not one specific aggregation rule), for example by an active approach that presents the conflict to the human and asks how it should be resolved, and learning an aggregation rule based on that. I'd be surprised if we ended up using a particular mathematical equation presented here as an aggregation mechanism -- I'm much more interested in what problems arise when we try to aggregate things, what criteria we might want to satisfy, etc.

Interpretability

[Towards Governing Agent's Efficacy: Action-Conditional \$\beta\$ -VAE for Deep Transparent Reinforcement Learning](#) (*John Yang et al*)

Verification

[Evaluating Robustness of Neural Networks with Mixed Integer Programming](#) (*Anonymous*): I've only read the abstract so far, but this paper claims to find the exact adversarial accuracy of an MNIST classifier within an L infinity norm ball of radius 0.1, which would be a big step forward in the state of the art for verification.

[On a Formal Model of Safe and Scalable Self-driving Cars](#) (*Shai Shalev-Shwartz et al*)

Robustness

[ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness \(Anonymous\)](#) (summarized by Dan H): This paper empirically demonstrates the outsized influence of textures in classification. To address this, they apply style transfer to ImageNet images and train with this dataset. Although training networks on a specific corruption tends to provide robustness only to that specific corruption, stylized ImageNet images supposedly lead to generalization to new corruption types such as uniform noise and high-pass filters (but not blurs).

[Learning Robust Representations by Projecting Superficial Statistics Out \(Anonymous\)](#)

AI strategy and policy

[AI development incentive gradients are not uniformly terrible \(rk\)](#): This post considers a model of AI development somewhat similar to the one in [Racing to the precipice](#) paper. It notes that under this model, assuming perfect information, the utility curves for each player are *discontinuous*. Specifically, the models predict deterministically that the player that spent the most on something (typically AI capabilities) is the one that "wins" the race (i.e. builds AGI), and so there is a discontinuity at the point where the players are spending equal amounts of money. This results in players fighting as hard as possible to be on the right side of the discontinuity, which suggests that they will skimp on safety. However, in practice, there will be some uncertainty about which player wins, even if you know exactly how much each is spending, and this removes the discontinuity. The resulting model predicts more investment in safety, since buying expected utility through safety now looks better than increasing the probability of winning the race (whereas before, it was compared against changing from definitely losing the race to definitely winning the race).

Rohin's opinion: The model in [Racing to the precipice](#) had the unintuitive conclusion that if teams have *more* information (i.e. they know their own or other's capabilities), then we become *less* safe, which puzzled me for a while. Their explanation is that with maximal information, the top team takes as much risk as necessary in order to guarantee that they beat the second team, which can be quite a lot of risk if the two teams are close. While this is true, the explanation from this post is more satisfying -- since the model has a discontinuity that rewards taking on risk, anything that removes the discontinuity and makes it more continuous will likely improve the prospects for safety, such as not having full information. I claim that in reality these discontinuities mostly don't exist, since (1) we're uncertain about who will win and (2) we will probably have a multipolar scenario where even if you aren't first-to-market you can still capture a lot of value. This suggests that it likely isn't a problem for teams to have more information about each other on the margin.

That said, these models are still very simplistic, and I mainly try to derive qualitative conclusions from them that my intuition agrees with in hindsight.

Prerequisites: [Racing to the precipice: a model of artificial intelligence development](#)

Other progress in AI

Reinforcement learning

[Learning Latent Dynamics for Planning from Pixels](#) (*Danijar Hafner et al*) (summarized by Richard): The authors introduce PlaNet, an agent that learns an environment's dynamics from pixels and then chooses actions by planning in latent space. At each step, it searches for the best action sequence under its Recurrent State Space dynamics model, then executes the first action and replans. The authors note that having a model with both deterministic and stochastic transitions is critical to learning a good policy. They also use a technique called variational overshooting to train the model on multi-step predictions, by generalising the standard variational bound for one-step predictions. PlaNet approaches the performance of top model-free algorithms even when trained on 50x fewer episodes.

Richard's opinion: This paper seems like a step forward in addressing the instability of using learned models in RL. However, the extent to which it's introducing new contributions, as opposed to combining existing ideas, is a little unclear.

[Modular Architecture for StarCraft II with Deep Reinforcement Learning](#) (*Dennis Lee, Haoran Tang et al*)

Deep learning

[Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet](#) (*Anonymous*) (summarized by Dan H): This paper proposes a bag-of-features model using patches as features, and they show that this can obtain accuracy similar to VGGNet architectures. They classify each patch and produce the final classification by a majority vote; Figure 1 of the paper tells all. In some ways this model is more interpretable than other deep architectures, as it is clear which regions activated which class. They attempt to show that, like their model, VGGNet does not use global shape information but instead uses localized features.

Machine learning

[Formal Limitations on The Measurement of Mutual Information](#) (*David McAllester and Karl Stratos*)

Humans can be assigned any values whatsoever...

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

(Re)Posted as part of the AI Alignment Forum sequence on [Value Learning](#).

Rohin's note: In the last [post](#), we saw that a good broad value learning approach would need to understand the systematic biases in human planning in order to achieve superhuman performance. Perhaps we can just use machine learning again and learn the biases and reward simultaneously? This post by Stuart Armstrong (original [here](#)) and the associated [paper](#) say: "Not without more assumptions."

This post comes from a theoretical perspective that may be alien to ML researchers; in particular, it makes an argument that simplicity priors do not solve the problem pointed out here, where simplicity is based on [Kolmogorov complexity](#) (which is an instantiation of the [Minimum Description Length principle](#)). The analog in machine learning would be an argument that regularization would not work. The proof used is specific to Kolmogorov complexity and does not clearly generalize to arbitrary regularization techniques; however, I view the argument as being suggestive that regularization techniques would also be insufficient to address the problems raised here.

Humans have no values... nor do any agent. Unless you make strong assumptions about their rationality. And depending on those assumptions, you get humans to have any values.

An agent with no clear preferences

There are three buttons in this world, B(0), B(1), and X, and one agent H.

B(0) and B(1) can be operated by H, while X can be operated by an outside observer. H will initially press button B(0); if ever X is pressed, the agent will switch to pressing B(1). If X is pressed again, the agent will switch back to pressing B(0), and so on. After a large number of turns N, H will shut off. That's the full algorithm for H.

So the question is, what are the values/preferences/rewards of H? There are three natural reward functions that are plausible:

- R(0), which is linear in the number of times B(0) is pressed.

- $R(1)$, which is linear in the number of times $B(1)$ is pressed.
- $R(2) = I(E, X)R(0) + I(O, X)R(1)$, where $I(E, X)$ is the indicator function for X being pressed an even number of times, $I(O, X) = 1 - I(E, X)$ being the indicator function for X being pressed an odd number of times.

For $R(0)$, we can interpret H as an $R(0)$ maximising agent which X overrides. For $R(1)$, we can interpret H as an $R(1)$ maximising agent which X releases from constraints. And $R(2)$ is the “ H is always fully rational” reward. Semantically, these make sense for the various $R(i)$ ’s being a true and natural reward, with X = “coercive brain surgery” in the first case, X = “release H from annoying social obligations” in the second, and X = “switch which of $R(0)$ and $R(1)$ gives you pleasure” in the last case.

But note that there is no semantic implications here, all that we know is H , with its full algorithm. If we wanted to deduce its true reward for the purpose of something like [Inverse Reinforcement Learning](#) (IRL), what would it be?

Modelling human (ir)rationality and reward

Now let’s talk about the preferences of an actual human. We all know that humans are not always rational. But even if humans were fully rational, the fact remains that we are physical, and vulnerable to things like coercive brain surgery (and in practice, to a whole host of other more or less manipulative techniques). So there will be the equivalent of “button X ” that overrides human preferences. Thus, “not immortal and unchangeable” is in practice enough for the agent to be considered “not fully rational”.

Now assume that we’ve thoroughly observed a given human h (including their internal brain wiring), so we know the human policy $\pi(h)$ (which determines their actions in all circumstances). This is, in practice all that we can ever observe - once we know $\pi(h)$ perfectly, there is nothing more that observing h can teach us.

Let R be a possible human reward function, and \mathbf{R} the set of such rewards. A human (ir)rationality planning algorithm p (hereafter referred to as a planner), is a map from \mathbf{R} to the space of policies (thus $p(R)$ says how a human with reward R will actually behave - for example, this could be bounded rationality, rationality with biases, or

many other options). Say that the pair (p, R) is compatible if $p(R) = \pi(h)$. Thus a human with planner p and reward R would behave as h does.

What possible compatible pairs are there? Here are some candidates:

- $(p(0), R(0))$, where $p(0)$ and $R(0)$ are some “plausible” or “acceptable” planner and reward functions (what this means is a big question).
- $(p(1), R(1))$, where $p(1)$ is the “fully rational” planner, and $R(1)$ is a reward that fits to give the required policy.
- $(p(2), R(2))$, where $R(2) = -R(1)$, and $p(2) = -p(1)$, where $-p(R)$ is defined as $p(-R)$; here $p(2)$ is the “fully anti-rational” planner.
- $(p(3), R(3))$, where $p(3)$ maps all rewards to $\pi(h)$, and $R(3)$ is trivial and constant.
- $(p(4), R(4))$, where $p(4) = -p(0)$ and $R(4) = -R(0)$.

Distinguishing among compatible pairs

How can we distinguish between compatible pairs? At first appearance, we can't. That's because, by their definition of compatible, all pairs produce the correct policy $\pi(h)$. And once we have $\pi(h)$, further observations of h tell us nothing.

I initially thought that Kolmogorov or algorithmic complexity might help us here. But in fact:

Theorem: The pairs $(p(i), R(i))$, $i \geq 1$, are either simpler than $(p(0), R(0))$, or differ in Kolmogorov complexity from it by a constant that is independent of $(p(0), R(0))$.

Proof: The cases of $i = 4$ and $i = 2$ are easy, as these differ from $i = 0$ and $i = 1$ by two minus signs. Given $(p(0), R(0))$, a fixed-length algorithm computes $\pi(h)$. Then a fixed length algorithm defines $p(3)$ (by mapping input to $\pi(h)$). Furthermore, given $\pi(h)$ and any history η , a fixed length algorithm computes the action $a(\eta)$ the agent will take; then a fixed length algorithm defines $R(1)(\eta, a(\eta)) = 1$ and $R(1)(\eta, b) = 0$ for $b \neq a(\eta)$.

So the Kolmogorov complexity can shift between p and R (all in R for $i = 1, 2$, all in p for $i = 3$), but it seems that the complexity of the pair doesn't go up during these

shifts.

This is puzzling. It seems that, in principle, one cannot assume anything about H's reward at all! $R(2) = -R(1)$, $R(4) = -R(0)$, and $p(3)$ is compatible with any possible reward R. If we give up the assumption of human rationality - which we must - it seems we can't say anything about the human reward function. So it seems IRL must fail.