

# Best of LessWrong: June 2022

1. [Where I agree and disagree with Eliezer](#)
2. [AGI Ruin: A List of Lethalities](#)
3. [It's Probably Not Lithium](#)
4. [Security Mindset: Lessons from 20+ years of Software Security Failures Relevant to AGI Alignment](#)
5. [What Are You Tracking In Your Head?](#)
6. [Humans are very reliable agents](#)
7. [A central AI alignment problem: capabilities generalization, and the sharp left turn](#)
8. [Contra Hofstadter on GPT-3 Nonsense](#)
9. [Slow motion videos as AI risk intuition pumps](#)
10. [AGI Safety FAQ / all-dumb-questions-allowed thread](#)
11. [Announcing the Inverse Scaling Prize \(\\$250k Prize Pool\)](#)
12. [The inordinately slow spread of good AGI conversations in ML](#)
13. [AI Could Defeat All Of Us Combined](#)
14. [Nonprofit Boards are Weird](#)
15. [On A List of Lethalities](#)
16. [Why all the fuss about recursive self-improvement?](#)
17. [LessWrong Has Agree/Disagree Voting On All New Comment Threads](#)
18. [The prototypical catastrophic AI action is getting root access to its datacenter](#)
19. [AI-Written Critiques Help Humans Notice Flaws](#)
20. [Godzilla Strategies](#)
21. [Public beliefs vs. Private beliefs](#)
22. [A transparency and interpretability tech tree](#)
23. [Announcing the LessWrong Curated Podcast](#)
24. [Conversation with Eliezer: What do you want the system to do?](#)
25. [Confused why a "capabilities research is good for alignment progress" position isn't discussed more](#)
26. [Contra EY: Can AGI destroy us without trial & error?](#)
27. [Intergenerational trauma impeding cooperative existential safety efforts](#)
28. [why assume AGIs will optimize for fixed goals?](#)
29. [A descriptive, not prescriptive, overview of current AI Alignment Research](#)
30. [Limits to Legibility](#)
31. [Let's See You Write That Corrigibility Tag](#)
32. [Leaving Google, Joining the Nucleic Acid Observatory](#)
33. ["Pivotal Acts" means something specific](#)
34. [Will Capabilities Generalise More?](#)
35. [CFAR Handbook: Introduction](#)
36. [Who models the models that model models? An exploration of GPT-3's in-context model fitting ability](#)
37. [Deep Learning Systems Are Not Less Interpretable Than Logic/Probability/Etc](#)
38. [wrapper-minds are the enemy](#)
39. [Contest: An Alien Message](#)
40. [Relationship Advice Repository](#)
41. [Yes, AI research will be substantially curtailed if a lab causes a major disaster](#)
42. [Units of Exchange](#)
43. [In defense of flailing, with foreword by Bill Burr](#)
44. [Air Conditioner Repair](#)
45. [Pivotal outcomes and pivotal processes](#)
46. [Air Conditioner Test Results & Discussion](#)
47. [Causal confusion as an argument against the scaling hypothesis](#)
48. [I'm trying out "asteroid mindset"](#)
49. [AI Training Should Allow Opt-Out](#)
50. [A Quick List of Some Problems in AI Alignment As A Field](#)

# Best of LessWrong: June 2022

1. [Where I agree and disagree with Eliezer](#)
2. [AGI Ruin: A List of Lethalities](#)
3. [It's Probably Not Lithium](#)
4. [Security Mindset: Lessons from 20+ years of Software Security Failures Relevant to AGI Alignment](#)
5. [What Are You Tracking In Your Head?](#)
6. [Humans are very reliable agents](#)
7. [A central AI alignment problem: capabilities generalization, and the sharp left turn](#)
8. [Contra Hofstadter on GPT-3 Nonsense](#)
9. [Slow motion videos as AI risk intuition pumps](#)
10. [AGI Safety FAQ / all-dumb-questions-allowed thread](#)
11. [Announcing the Inverse Scaling Prize \(\\$250k Prize Pool\)](#)
12. [The inordinately slow spread of good AGI conversations in ML](#)
13. [AI Could Defeat All Of Us Combined](#)
14. [Nonprofit Boards are Weird](#)
15. [On A List of Lethalities](#)
16. [Why all the fuss about recursive self-improvement?](#)
17. [LessWrong Has Agree/Disagree Voting On All New Comment Threads](#)
18. [The prototypical catastrophic AI action is getting root access to its datacenter](#)
19. [AI-Written Critiques Help Humans Notice Flaws](#)
20. [Godzilla Strategies](#)
21. [Public beliefs vs. Private beliefs](#)
22. [A transparency and interpretability tech tree](#)
23. [Announcing the LessWrong Curated Podcast](#)
24. [Conversation with Eliezer: What do you want the system to do?](#)
25. [Confused why a "capabilities research is good for alignment progress" position isn't discussed more](#)
26. [Contra EY: Can AGI destroy us without trial & error?](#)
27. [Intergenerational trauma impeding cooperative existential safety efforts](#)
28. [why assume AGIs will optimize for fixed goals?](#)
29. [A descriptive, not prescriptive, overview of current AI Alignment Research](#)
30. [Limits to Legibility](#)
31. [Let's See You Write That Corrigibility Tag](#)
32. [Leaving Google, Joining the Nucleic Acid Observatory](#)
33. ["Pivotal Acts" means something specific](#)
34. [Will Capabilities Generalise More?](#)
35. [CFAR Handbook: Introduction](#)
36. [Who models the models that model models? An exploration of GPT-3's in-context model fitting ability](#)
37. [Deep Learning Systems Are Not Less Interpretable Than Logic/Probability/Etc](#)
38. [wrapper-minds are the enemy](#)
39. [Contest: An Alien Message](#)
40. [Relationship Advice Repository](#)
41. [Yes, AI research will be substantially curtailed if a lab causes a major disaster](#)
42. [Units of Exchange](#)
43. [In defense of flailing, with foreword by Bill Burr](#)
44. [Air Conditioner Repair](#)
45. [Pivotal outcomes and pivotal processes](#)

46. [Air Conditioner Test Results & Discussion](#)
47. [Causal confusion as an argument against the scaling hypothesis](#)
48. [I'm trying out "asteroid mindset"](#)
49. [AI Training Should Allow Opt-Out](#)
50. [A Quick List of Some Problems in AI Alignment As A Field](#)

# Where I agree and disagree with Eliezer

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

(Partially in response to [AGI Ruin: A list of Lethalities](#). Written in the same rambling style. Not exhaustive.)

## Agreements

1. Powerful AI systems have a good chance of deliberately and irreversibly disempowering humanity. This is a much easier failure mode than killing everyone with destructive physical technologies.
2. Catastrophically risky AI systems could plausibly exist soon, and there likely won't be a strong consensus about this fact until such systems pose a meaningful existential risk per year. There is not necessarily any "fire alarm."
3. Even if there were consensus about a risk from powerful AI systems, there is a good chance that the world would respond in a totally unproductive way. It's wishful thinking to look at possible stories of doom and say "we wouldn't let that happen;" humanity is fully capable of messing up even very basic challenges, especially if they are novel.
4. I think that many of the projects intended to help with AI alignment don't make progress on key difficulties and won't significantly reduce the risk of catastrophic outcomes. This is related to people gravitating to whatever research is most tractable and not being too picky about what problems it helps with, and related to a low level of concern with the long-term future in particular. Overall, there are relatively few researchers who are effectively focused on the technical problems most relevant to existential risk from alignment failures.
5. There are strong social and political pressures to spend much more of our time talking about how AI shapes existing conflicts and shifts power. This pressure is already playing out and it doesn't seem too likely to get better. I think Eliezer's term "[the last derail](#)" is hyperbolic but on point.
6. Even when thinking about accident risk, people's minds seem to go to what they think of as "more realistic and less sci fi" risks that are much less likely to be existential (and sometimes I think less plausible). It's very possible this dynamic won't change until after actually existing AI systems pose an existential risk.
7. There is a good chance that an AI catastrophe looks like an abrupt "coup" where AI systems permanently disempower humans with little opportunity for resistance. People seem to consistently round this risk down to more boring stories that fit better with their narratives about the world. It is quite possible that an AI coup will be sped up by humans letting AI systems control killer robots, but the difference in timeline between "killer robots everywhere, AI controls everything" and "AI only involved in R&D" seems like it's less than a year.
8. The broader intellectual world seems to wildly overestimate how long it will take AI systems to go from "large impact on the world" to "unrecognizably transformed world." This is more likely to be years than decades, and there's a real chance that it's months. This makes alignment harder and doesn't seem like something we are collectively prepared for.

9. Humanity usually solves technical problems by iterating and fixing failures; we often resolve tough methodological disagreements very slowly by seeing what actually works and having our failures thrown in our face. But it will probably be possible to build valuable AI products without solving alignment, and so reality won't "force us" to solve alignment until it's too late. This seems like a case where we will have to be unusually reliant on careful reasoning rather than empirical feedback loops for some of the highest-level questions.
10. AI systems will ultimately be wildly superhuman, and there probably won't be strong technological hurdles right around human level. Extrapolating the rate of existing AI progress suggests you don't get too much time between weak AI systems and very strong AI systems, and AI contributions could very easily go from being a tiny minority of intellectual work to a large majority over a few years.
11. If you had incredibly powerful unaligned AI systems running on a server farm somewhere, there is very little chance that humanity would maintain meaningful control over its future.
12. "Don't build powerful AI systems" appears to be a difficult policy problem, requiring geopolitical coordination of a kind that has often failed even when the stakes are unambiguous and the pressures to defect are much smaller.
13. I would not expect humanity to necessarily "rise to the challenge" when the stakes of a novel problem are very large. I was 50-50 about this in 2019, but our experience with COVID has further lowered my confidence.
14. There is probably no physically-implemented reward function, of the kind that could be optimized with SGD, that we'd be happy for an arbitrarily smart AI to optimize as hard as possible. (I'm most optimistic about approaches where RL is only performed on a reward function that gets smarter in parallel with the agent being trained.)
15. Training an AI to maximize a given reward function does not generically produce an AI which is internally "motivated" to maximize reward. Moreover, at some level of capability, a very wide range of motivations for an AI would lead to loss-minimizing behavior on the training distribution because minimizing loss is an important strategy for an AI to preserve its influence over the world.
16. It is more robust for an AI system to learn a good model for the environment, and what the consequences of its actions will be, than to learn a behavior like "generally being nice" or "trying to help humans." Even if an AI was imitating data consisting of "what I would do if I were trying to be nice," it would *still* be more likely to eventually learn to imitate the actual physical process producing that data rather than absorbing some general habit of niceness. And in practice the data we produce will not be perfect, and so "predict the physical process generating your losses" is going to be positively selected for by SGD.
17. You shouldn't say something like "well I might as well assume there's a hope" and thereby live in a specific unlikely world where alignment is unrealistically easy in one way or another. Even if alignment ends up easy, you would be likely to end up predicting the wrong way for it to be easy. If things look doomed to you, in practice it's better to try to maximize log odds of success as a more general and robust strategy for taking advantage of lucky breaks in a messy and hard-to-predict world.
18. No current plans for aligning AI have a particularly high probability of working without a lot of iteration and modification. The current state of affairs is roughly "if alignment turns out to be a real problem, we'll learn a lot about it and iteratively improve our approach." If the problem is severe and emerges quickly, it would be better if we had a clearer plan further in advance—we'd still have to adapt and learn, but starting with something that looks like it could work on paper would put us in a much better situation.

19. Many research problems in other areas are chosen for tractability or being just barely out of reach. We pick benchmarks we can make progress on, or work on theoretical problems that seem well-posed and approachable using existing techniques. Alignment isn't like that; it was chosen to be an important problem, and there is no one ensuring that the game is "fair" and that the problem is soluble or tractable.

## Disagreements

(*Mostly stated without argument.*)

1. Eliezer often equivocates between "you have to get alignment right on the first 'critical' try" and "you can't learn anything about alignment from experimentation and failures before the critical try." This distinction is very important, and I agree with the former but disagree with the latter. Solving a scientific problem without being able to learn from experiments and failures is incredibly hard. But we will be able to learn a lot about alignment from experiments and trial and error; I think we can get a lot of feedback about what works and deploy more traditional R&D methodology. We have toy models of alignment failures, we have standards for interpretability that we can't yet meet, and we have theoretical questions we can't yet answer.. The difference is that reality doesn't force us to solve the problem, or tell us clearly which analogies are the right ones, and so it's possible for us to push ahead and build AGI without solving alignment. Overall this consideration seems like it makes the *institutional* problem vastly harder, but does not have such a large effect on the scientific problem.
2. Eliezer often talks about AI systems that are able to easily build nanotech and overpower humans decisively, and describes a vision of a rapidly unfolding doom from a single failure. This is what would happen if you were magically given an extraordinarily powerful AI and then failed to aligned it, but I think it's very unlikely what will happen in the real world. By the time we have AI systems that can overpower humans decisively with nanotech, we have other AI systems that will either kill humans in more boring ways or else radically advanced the state of human R&D. More generally, the cinematic universe of Eliezer's stories of doom doesn't seem to me like it holds together, and I can't tell if there is a more realistic picture of AI development under the surface.
3. One important factor seems to be that Eliezer often imagines scenarios in which AI systems avoid making major technical contributions, or revealing the extent of their capabilities, because they are lying in wait to cause trouble later. But if we are constantly training AI systems to do things that look impressive, then SGD will be aggressively selecting against any AI systems who don't do impressive-looking stuff. So by the time we have AI systems who can develop molecular nanotech, we will definitely have had systems that did something slightly-less-impressive-looking.
4. AI improving itself is most likely to look like AI systems doing R&D in the same way that humans do. "AI smart enough to improve itself" is not a crucial threshold, AI systems will get gradually better at improving themselves. Eliezer appears to expect AI systems performing extremely fast recursive self-improvement before those systems are able to make superhuman contributions to other domains (including alignment research), but I think this is mostly unjustified. If Eliezer doesn't believe this, then his arguments about the alignment problem that *humans need to solve* appear to be wrong.

5. The notion of an AI-enabled “pivotal act” seems misguided. Aligned AI systems can reduce the period of risk of an unaligned AI by advancing alignment research, convincingly demonstrating the risk posed by unaligned AI, and consuming the “[free energy](#)” that an unaligned AI might have used to grow explosively. No particular act needs to be pivotal in order to greatly reduce the risk from unaligned AI, and the search for single pivotal acts leads to unrealistic stories of the future and unrealistic pictures of what AI labs should do.
6. Many of the “pivotal acts” that Eliezer discusses involve an AI lab achieving a “decisive strategic advantage” (i.e. overwhelming hard power) that they use to implement a relatively limited policy, e.g. restricting the availability of powerful computers. But the same hard power would also let them arbitrarily dictate a new world order, and would be correctly perceived as an existential threat to existing states. Eliezer’s view appears to be that a decisive strategic advantage is the most realistic way to achieve these policy goals, *despite* the fact that building powerful enough AI systems runs an overwhelming risk of destroying the world via misalignment. I think that preferring this route to more traditional policy influence requires extreme confidence about details of the policy situation; that confidence might be justified by someone who knew a lot more about the details of government than I do, but Eliezer does not seem to. While I agree that this kind of policy change would be an unusual success in historical terms, the probability still seems much higher than Eliezer’s overall probabilities of survival. Conversely, I think Eliezer greatly underestimates how difficult it would be for an AI developer to covertly take over the world, how strongly and effectively governments would respond to that possibility, and how toxic this kind of plan is.
7. I think Eliezer is probably wrong about how useful AI systems will become, including for tasks like AI alignment, before it is catastrophically dangerous. I believe we are relatively quickly approaching AI systems that can meaningfully accelerate progress by generating ideas, recognizing problems for those ideas and, proposing modifications to proposals, etc. and that all of those things will become possible in a small way well before AI systems that can double the pace of AI research. By the time AI systems can double the pace of AI research, it seems like they can greatly accelerate the pace of alignment research. Eliezer is right that this doesn’t make the problem go away (if humans don’t solve alignment, then why think AIs will solve it?) but I think it does mean that arguments about how recursive self-improvement quickly kicks you into a lethal regime are wrong (since AI is accelerating the timetable for both alignment and capabilities).
8. When talking about generalization outside of the training distribution, I think Eliezer is generally pretty sloppy. I think many of the points are roughly right, but that it’s way too sloppy to reach reasonable conclusions after several steps of inference. I would love to see real discussion of these arguments, and in some sense it seems like Eliezer is a good person to push that discussion forward. Right now I think that relevant questions about ML generalization are in fact pretty subtle; we can learn a lot about them in advance but right now just mostly don’t know. Similarly, I think Eliezer’s reasoning about convergent incentives and the deep nature of consequentialism is too sloppy to get to correct conclusions and the resulting assertions are wildly overconfident.
9. In particular, existing AI training strategies don’t need to handle a “drastic” distribution shift from low levels of intelligence to high levels of intelligence. There’s nothing in the foreseeable ways of building AI that would call for a big transfer like this, rather than continuously training as intelligence gradually increases. Eliezer seems to partly be making a relatively confident claim that the nature of AI is going to change a lot, which I think is probably wrong and is

clearly overconfident. If he had been actually making concrete predictions over the last 10 years I think he would be losing a lot of them to people more like me.

10. Eliezer strongly expects sharp capability gains, based on a combination of arguments that I think don't make sense and an analogy with primate evolution which I think is being applied poorly. We've talked about this before, and I still think Eliezer's position is probably wrong and clearly overconfident. I find Eliezer's more detailed claims, e.g. about hard thresholds, to be much more implausible than his (already probably quantitatively wrong) claims about takeoff speeds.
11. Eliezer seems confident about the difficulty of alignment based largely on his own experiences working on the problem. But in fact society has spent very little total effort working on the problem, and MIRI itself would probably be unable to solve or even make significant progress on the large majority of problems that existing research fields routinely solve. So I think right now we mostly don't know how hard the problem is (but it may well be very hard, and even if it's easy we may well fail to solve it). For example, the fact that MIRI tried and failed to find a "coherent formula for corrigibility" is not much evidence that corrigibility is "unworkable."
12. Eliezer says a lot of concrete things about how research works and about what kind of expectation of progress is unrealistic (e.g. talking about bright-eyed optimism in [list of lethali](#)ties). But I don't think this is grounded in an understanding of the history of science, familiarity with the dynamics of modern functional academic disciplines, or research experience. The Eliezer predictions most relevant to "how do scientific disciplines work" that I'm most aware of are incorrectly predicting that physicists would be wrong about the existence of the Higgs boson ([LW bet registry](#)) and expressing the view that real AI would likely emerge from a small group rather than a large industry ([pg.436](#) but expressed many places).
13. I think Eliezer generalizes a lot from pessimism about solving problems easily to pessimism about solving problems at all; or from the fact that a particular technique doesn't immediately solve a problem to pessimism about the helpfulness of research on that technique. I disagree with Eliezer about how research progress is made, and don't think he has any special expertise on this topic. Eliezer often makes objections to particular implementations of projects (like using interpretability tools for training). But in order to actually talk about whether a research project is likely to succeed, you really really need to engage with the existential quantifier where future researchers get to choose implementation details to make it work. At a *minimum* that requires engaging with the strongest existing versions of these proposals, and if you haven't done that (as Eliezer hasn't) then you need to take a different kind of approach. But even if you engage with the best existing concrete proposals, you still need to think carefully about whether your objections are the kind of thing that will be hard to overcome as people learn more details in the future. One way of looking at this is that Eliezer is appropriately open-minded about existential quantifiers applied to future AI systems thinking about how to cause trouble, but seems to treat existential quantifiers applied to future humans in a qualitatively rather than quantitatively different way (and as described throughout this list I think he overestimates the quantitative difference).
14. As an example, I think Eliezer is unreasonably pessimistic about interpretability while being mostly ignorant about the current state of the field. This is true both for the level of understanding potentially achievable by interpretability, and the possible applications of such understanding. I agree with Eliezer that this seems like a hard problem and many people seem unreasonably optimistic, so I might be sympathetic if Eliezer was making claims with moderate confidence rather

than high confidence. As far as I can tell most of Eliezer's position here comes from general intuitions rather than arguments, and I think those are much less persuasive when you don't have much familiarity with the domain.

15. Early transformative AI systems will probably do impressive technological projects by being trained on smaller tasks with shorter feedback loops and then composing these abilities in the context of large collaborative projects (initially involving a lot of humans but over time increasingly automated). When Eliezer dismisses the possibility of AI systems performing safer tasks millions of times in training and then safely transferring to "build nanotechnology" (point 11 of [list of lethali](#)[ties](#)) he is not engaging with the kind of system that is likely to be built or the kind of hope people have in mind.
16. [List of lethali](#)[ties](#) #13 makes a particular argument that we won't see many AI problems in advance; I feel like I see this kind of thinking from Eliezer a lot but it seems misleading or wrong. In particular, it seems possible to study the problem that AIs may "change [their] outer behavior to deliberately look more aligned and deceive the programmers, operators, and possibly any loss functions optimizing over [them]" in advance. And while it's true that *if you fail to solve that problem then you won't notice other problems*, this doesn't really affect the probability of solving alignment overall: if you don't solve that problem then you die, and if you do solve that problem then you can study the other problems.
17. I don't think [list of lethali](#)[ties](#) is engaging meaningfully with the most serious hopes about how to solve the alignment problem. I don't think that's necessarily the purpose of the list, but it's quite important if you want to assess the probability of doom or contribute meaningfully to solving the problem (or to complain about other people producing similar lists).
18. I think that natural selection is a relatively weak analogy for ML training. The most important disanalogy is that we can deliberately shape ML training. Animal breeding would be a better analogy, and seems to suggest a different and much more tentative conclusion. For example, if humans were being actively bred for corrigibility and friendliness, it looks to me like they would quite likely be corrigible and friendly up through the current distribution of human behavior. If that breeding process was continuously being run carefully by the smartest of the currently-friendly humans, it seems like it would plausibly break down at a level very far beyond current human abilities.
19. Eliezer seems to argue that humans couldn't verify pivotal acts proposed by AI systems (e.g. contributions to alignment research), and that this further makes it difficult to safely perform pivotal acts. In addition to disliking his concept of pivotal acts, I think that this claim is probably wrong and clearly overconfident. I think it doesn't match well with pragmatic experience in R&D in almost any domain, where verification is *much, much* easier than generation in virtually every domain.
20. Eliezer is relatively confident that you can't train powerful systems by imitating human thoughts, because too much of human thinking happens under the surface. I think this is fairly plausible but it's not at all obvious, and moreover there are plenty of techniques intermediate between "copy individual reasoning steps" and "optimize end-to-end on outcomes." I think that the last 5 years of progress in language modeling have provided significant evidence that training AI to imitate human thought may be economically competitive at the time of transformative AI, potentially bringing us to something more like a 50-50 chance. I can't tell if Eliezer should have lost Bayes points here, but I suspect he would have and if he wants us to evaluate his actual predictions I wish he would say *something* about his future predictions.

21. These last two points (and most others from this list) aren't actually part of my central alignment hopes or plans. Alignment hopes, like alignment concerns, can be disjunctive. In some sense they are even more disjunctive, since the existence of humans who are trying to solve alignment is considerably more robust than the existence of AI systems who are trying to cause trouble (such AIs only exist if humans have *already failed* at significant parts of alignment). Although my research is focused on cases where almost every factor works out against us, I think that you can get a lot of survival probability from easier worlds.
22. Eliezer seems to be relatively confident that AI systems will be very alien and will understand many things about the world that humans don't, rather than understanding a similar profile of things (but slightly better), or having weaker understanding but enjoying other advantages like much higher serial speed. I think this is very unclear and Eliezer is wildly overconfident. It seems plausible that AI systems will learn much of how to think by predicting humans even *if* human language is a uselessly shallow shadow of human thought, because of the extremely short feedback loops. It also seems quite possible that most of their knowledge about science will be built by an explicit process of scientific reasoning and inquiry that will proceed in a recognizable way to human science even if their minds are quite different. Most importantly, it seems like AI systems have huge structural advantages (like their high speed and low cost) that suggest they will have a transformative impact on the world (~~and obsolete human contributions to alignment~~ [retracted](#)) well before they need to develop superhuman understanding of much of the world or tricks about how to think, and so even if they have a very different profile of abilities to humans they may still be subhuman in many important ways.
23. AI systems reasoning about the code of other AI systems is not likely to be an important dynamic for early cooperation between AIs. Those AI systems look very likely to be messy, such that the only way AI systems will reason about their own or others' code is by looking at behavior and using the same kinds of tools and reasoning strategies as humans. Eliezer has a consistent pattern of identifying important long-run considerations, and then flatly asserting that they are relevant in the short term without evidence or argument. I think Eliezer thinks this pattern of predictions isn't yet conflicting with the evidence because these predictions only kick in at some later point (but still early enough to be relevant), but this is part of what makes his prediction track record impossible to assess and why I think he is greatly overestimating it in hindsight.
24. Eliezer's model of AI systems cooperating with each other to undermine "checks and balances" seems wrong to me, because it focuses on cooperation and the incentives of AI systems. Realistic proposals mostly don't need to rely on the *incentives* of AI systems, they can instead rely on gradient descent *selecting* for systems that play games competitively, e.g. by searching until we find an AI which raises compelling objections to other AI systems' proposals. (Note: I may be misunderstanding what he's saying in places like [list of lethali](#)ties 35; those may be a response to random things people say on the internet rather than engaging with alignment strategies that are being explored in practice. But even then I think he is responding to relatively weak versions of these arguments.) Eliezer equivocates between a line like "AI systems will cooperate" and "The verifiable activities you could use gradient descent to select for won't function appropriately as checks and balances." But Eliezer's position is a conjunction that fails if either step fails, and jumping back and forth between them appears to totally obscure the actual structure of the argument.
25. Eliezer seems to expect that by the time AI systems can achieve pivotal acts, they can also manipulate humans very well—such that for example it's pointless

to try to have debates between them or to try to play an adversarial game against them. But if we imagine minds with a human profile of abilities, it seems like they would be superhuman at R&D around the same time they were superhuman at persuasion, and could easily go either way. Moreover, there are a lot of reasons the AI seems *much* more likely to be superhuman at R&D if that's what we want: it's trained primarily to work on R&D, it's using tools and structure designed to facilitate R&D, there is a large collaboration amongst AIs to advance R&D whereas manipulation is often being performed by individual AI systems trying to win a debate. And on top of all that, if anything I very weakly expect AI systems' comparative advantage to be in R&D rather than human manipulation (since R&D is wildly out of distribution for humans).

26. I don't think surviving worlds have a plan in the sense Eliezer is looking for. Based on what Eliezer says I don't feel like he has a clear or accurate picture of what successful "plans" look like in the real world. I don't see any particular reason to defer to Eliezer at all on this point.
27. Eliezer says that his [list of lethali](#)ties is the kind of document that other people couldn't write and therefore shows they are unlikely to contribute (point 41). I think that's wrong. I think Eliezer's document is mostly aimed at rhetoric or pedagogy rather than being a particularly helpful contribution to the field that others should be expected to have prioritized; I think that which ideas are "important" is mostly a consequence of Eliezer's idiosyncratic intellectual focus rather than an objective fact about what is important; the main contributions are collecting up points that have been made in the past and ranting about them and so they mostly reflect on Eliezer-as-writer; and perhaps most importantly, I think more careful arguments on more important difficulties are in fact being made in other places. For example, ARC's [report on ELK](#) describes at least 10 difficulties of the same type and severity as the ~20 technical difficulties raised in Eliezer's list. About half of them are overlaps, and I think the other half are if anything more important since they are more relevant to core problems with realistic alignment strategies.<sup>[1]</sup>

## My take on Eliezer's takes

- Eliezer raises many good considerations backed by pretty clear arguments, but makes confident assertions that are much stronger than anything suggested by actual argument.
- Eliezer's post (and most of his writing) isn't bringing much new evidence to the table; it mostly either reasons *a priori* or draws controversial conclusions from uncontroversial evidence. I think that calls for a different approach than Eliezer has taken historically (if the goal was to productively resolve these disagreements).
  - I think that these arguments mostly haven't been written down publicly so that they can be examined carefully or subject to criticism. It's not clear whether Eliezer has the energy to do that, but I think that people who think that Eliezer's position is important should try to understand the arguments well enough to do that.
  - I think that people with Eliezer's views haven't engaged very much productively with people who disagree (and have often made such engagement hard). I think that if you really dive into any of these key points you will quickly reach details where Eliezer cannot easily defend his view to a smart disinterested audience. And I don't think that Eliezer could pass an ideological Turing test for people who disagree.

- I think those are valuable steps to take if you have a contrarian take of great importance, which remains controversial even within your weird corner of the world, and whose support comes almost entirely from reasoning and argument.
- A lot of the post seems to rest on intuitions and ways of thinking that Eliezer feels are empirically supported (rather than on arguments that can be explicitly stated). But I don't feel like I actually have much evidence about that, so I think it really does just come down to the arguments.
  - I think Eliezer would like to say that the last 20 years give a lot of evidence for his object-level intuitions and general way of thinking about the world. If that's the case, I think we should very strongly expect that he can state predictions about the future that will systematically be better than those of people who don't share his intuitions or reasoning strategies. I remain happy to make predictions about any questions he thinks would provide this kind of evidence, or to state a bunch of random questions where I'm happy to predict (where I think he will probably slightly underperform me). If there aren't any predictions about the future where these intuitions and methodologies overperform, I think you should be very skeptical that they got a lot of evidence over the last 20 years (and that's at least something that requires explanation).
  - I think Eliezer could develop good intuition about these topics that is "backed up" by predicting the results of more complicated arguments using more broadly-accepted reasoning principles. Similarly, a mathematician might have great intuitions about the truth of a theorem, and those intuitions could come entirely from feedback loops involving formal proofs rather than empirical data. But if two mathematicians had differing intuitions about a theorem, and their intuitions both came from formally proving a bunch of similar theorems, then I think the way to settle the disagreement is by using the normal rules of logic governing proofs. So this brings us back to the previous bullet point, and I think Eliezer should be more interested in actually making arguments and engaging with legitimate objections.
  - I don't think Eliezer has any kind of track record of exhibiting understanding in other ways (e.g. by accomplishing technological goals or other projects that require engaging with details of the world or making good day-to-day predictions). I think that's OK, but it means that I more strongly expect any empirically-backed intuitions to be cashed out as either predictions from afar or more careful arguments.

## 1. $\triangleleft$

Ten examples off the top of my head, that I think are about half overlapping and where I think the discussions in the ELK doc are if anything more thorough than the discussions in the list of lethaliites:

1. Goals defined in terms of sense data are manipulable by an AI who can compromise sensors, and this is a serious obstruction to using ML to optimize what we actually care about.
2. An AI may manipulate sensors by exploiting facts about the world or modes of heuristic reasoning that humans are totally unfamiliar with, such that humans couldn't recognize such tampering even if they spent a very long time examining proposed actions.
3. The human process of scientific understanding, even if automated, may end up being significantly less efficient than the use of gradient descent to

find opaque models of the world. In this case, it may be inevitable that AI systems understand things about the world we don't even if they try to help us do science.

4. If an AI is trained to predict human judgments or optimize scores as assessed by humans, then humans are likely to make errors. An AI system will eventually learn these errors rather than learning the intended behavior. Even if these errors aren't themselves important, it will then predictably copy human errors out of distribution leading to catastrophic outcomes.
5. Even if humans make no errors in the training set, an AI which understands the world already has a model of a human which can be quickly repurposed to make good predictions about human judgments, and so it will tend to do this and therefore copy human errors off distribution.
6. Even if the AI has no model of a human, in the limit where the AI's model is very complex and alien it is still faster and simpler for the AI to learn a model of "what a human would say" from scratch then to learn the intended ontology identification. So we can't count on SGD.
7. There are many training strategies that can train an AI to answer questions even in cases where humans could not answer correctly. However most of the approaches we know now, including those being explored in practice, seem to consistently top out at "questions that humans could answer if they have a lot more compute" which does not always seem good enough.
8. We could imagine more elaborate games where the easiest strategy for the AI is honesty, and then to regularize on computation time in order to learn an honest policy, but those require us to be careful about the construction of the training data in order ensure that the task is sufficiently hard, and there are no existing proposals that have that property. It's very hard to even set up games for which no strategy can outperform honesty.
9. Even if you were optimizing based on reliable observations of the real world, there are many bad actions that have no human-legible consequences for many years. At the point when legible consequences materialize it may be in a world that is too complex for existing humans to evaluate whether they are good or bad. If we don't build an AI that understands our preferences about this kind of subtle bad behavior, then a competitive world will push us into a bad outcome.
10. If the simplest policy to succeed at our task is a learned optimizer, and we try to regularize our AI to e.g. answer questions quickly, then its best strategy may be to internally searching for a policy which answers questions slowly (because it's quicker to find such a policy, and the time taken by the search is larger than the time taken by the mesapolicy). This makes it difficult to lean on regularization strategies to incentivize honesty.

# AGI Ruin: A List of Lethalities

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

## Preamble:

(If you're already familiar with all basics and don't want any preamble, skip ahead to [Section B](#) for technical difficulties of alignment proper.)

I have several times failed to write up a well-organized list of reasons why AGI will kill you. People come in with different ideas about why AGI would be survivable, and want to hear different *obviously* key points addressed first. Some fraction of those people are loudly upset with me if the obviously most important points aren't addressed immediately, and I address different points first instead.

Having failed to solve this problem in any good way, I now give up and solve it poorly with a poorly organized list of individual rants. I'm not particularly happy with this list; the alternative was publishing nothing, and publishing this seems marginally more [dignified](#).

Three points about the general subject matter of discussion here, numbered so as not to conflict with the list of lethalities:

**-3.** I'm assuming you are already familiar with some basics, and already know what '[orthogonality](#)' and '[instrumental convergence](#)' are and why they're true. People occasionally claim to me that I need to stop fighting old wars here, because, those people claim to me, those wars have already been won within the important-according-to-them parts of the current audience. I suppose it's at least true that none of the current major EA funders seem to be visibly in denial about orthogonality or instrumental convergence as such; so, fine. If you don't know what '[orthogonality](#)' or '[instrumental convergence](#)' are, or don't see for yourself why they're true, you need a different introduction than this one.

**-2.** When I say that alignment is lethally difficult, I am not talking about ideal or perfect goals of 'provable' alignment, nor total alignment of superintelligences on exact human values, nor getting AIs to produce satisfactory arguments about moral dilemmas which sorta-reasonable humans disagree about, nor attaining an absolute certainty of an AI not killing everyone. When I say that alignment is difficult, I mean that in practice, using the techniques we actually have, "please don't disassemble literally everyone with probability roughly 1" is an overly large ask that we are not on course to get. So far as I'm concerned, [if you can get a powerful AGI that carries out some pivotal superhuman engineering task, with a less than fifty percent chance of killing more than one billion people](#), I'll take it. Even smaller chances of killing even fewer people would be a nice luxury, but if you can get as incredibly far as "less than roughly certain to kill everybody", then you can probably get down to under a 5% chance with only slightly more effort. Practically all of the difficulty is in getting to "less than certainty of killing literally everyone". Trolley problems are not an interesting subproblem in all of this; if there are any survivors, you solved alignment. At this point, I no longer care how it works, I don't care how you got there, I am cause-agnostic about whatever methodology you used, all I am looking at is prospective results, all I want is that we have justifiable cause to believe of a pivotally useful AGI

'this will not kill literally everyone'. Anybody telling you I'm asking for stricter 'alignment' than this has failed at reading comprehension. The big ask from AGI alignment, the basic challenge I am saying is too difficult, is to obtain by any strategy whatsoever a significant chance of there being any survivors.

**-1.** None of this is about anything being impossible in principle. The metaphor I usually use is that if a textbook from one hundred years in the future fell into our hands, containing all of the simple ideas *that actually work robustly in practice*, we could probably build an aligned superintelligence in six months. For people schooled in machine learning, I use as my metaphor the difference between ReLU activations and sigmoid activations. Sigmoid activations are complicated and fragile, and do a terrible job of transmitting gradients through many layers; ReLUs are incredibly simple (for the unfamiliar, the activation function is literally  $\max(x, 0)$ ) and work much better. Most neural networks for the first decades of the field used sigmoids; the idea of ReLUs wasn't discovered, validated, and popularized until decades later. What's lethal is that we do not *have* the Textbook From The Future telling us all the simple solutions that actually in real life just work and are robust; we're going to be doing everything with metaphorical sigmoids on the first critical try. No difficulty discussed here about AGI alignment is claimed by me to be impossible - to merely human science and engineering, let alone in principle - if we had 100 years to solve it using unlimited retries, the way that science *usually* has an unbounded time budget and unlimited retries. This list of lethaliies is about things *we are not on course to solve in practice in time on the first critical try*; none of it is meant to make a much stronger claim about things that are *impossible in principle*.

That said:

Here, from my perspective, are some different true things that could be said, to contradict various false things that various different people seem to believe, about why AGI would be survivable on anything remotely resembling the current pathway, or any other pathway we can easily jump to.

## Section A:

This is a very lethal problem, it has to be solved one way or another, it has to be solved at a minimum strength and difficulty level instead of various easier modes that some dream about, we do not have any visible option of 'everyone' retreating to only solve safe weak problems instead, and failing on the first really dangerous try is fatal.

**1.** Alpha Zero blew past all accumulated human knowledge about Go after a day or so of self-play, with no reliance on human playbooks or sample games. Anyone relying on "well, it'll get up to human capability at Go, but then have a hard time getting past that because it won't be able to learn from humans any more" would have relied on vacuum. **AGI will not be upper-bounded by human ability or human learning speed. Things much smarter than human would be able to learn from less evidence than humans require** to have ideas driven into their brains; there are theoretical upper bounds here, but those upper bounds seem very high. (Eg, each bit of information that couldn't already be fully predicted can eliminate at most half the probability mass of all hypotheses under consideration.) It is not naturally (by default,

barring intervention) the case that everything takes place on a timescale that makes it easy for us to react.

**2. A cognitive system with sufficiently high cognitive powers, given any medium-bandwidth channel of causal influence, will not find it difficult to bootstrap to overpowering capabilities independent of human infrastructure.**

The concrete example I usually use here is nanotech, because there's been pretty detailed analysis of what definitely look like physically attainable lower bounds on what should be possible with nanotech, and those lower bounds are sufficient to carry the point. My lower-bound model of "how a sufficiently powerful intelligence would kill everyone, if it didn't want to not do that" is that it gets access to the Internet, emails some DNA sequences to any of the many many online firms that will take a DNA sequence in the email and ship you back proteins, and bribes/persuades some human who has no idea they're dealing with an AGI to mix proteins in a beaker, which then form a first-stage nanofactory which can build the actual nanomachinery. (Back when I was first deploying this visualization, the wise-sounding critics said "Ah, but how do you know even a superintelligence could solve the protein folding problem, if it didn't already have planet-sized supercomputers?" but one hears less of this after the advent of AlphaFold 2, for some odd reason.) The nanomachinery builds diamondoid bacteria, that replicate with solar power and atmospheric CHON, maybe aggregate into some miniature rockets or jets so they can ride the jetstream to spread across the Earth's atmosphere, get into human bloodstreams and hide, strike on a timer. **Losing a conflict with a high-powered cognitive system looks at least as deadly as "everybody on the face of the Earth suddenly falls over dead within the same second".** (I am using awkward constructions like 'high cognitive power' because standard English terms like 'smart' or 'intelligent' appear to me to function largely as status synonyms.

'Superintelligence' sounds to most people like 'something above the top of the status hierarchy that went to double college', and they don't understand why that would be all that dangerous? Earthlings have no word and indeed no standard native concept that means 'actually useful cognitive power'. A large amount of failure to panic sufficiently, seems to me to stem from a lack of appreciation for the incredible potential lethality of this thing that Earthlings as a culture have not named.)

**3. We need to get alignment right on the 'first critical try' at operating at a 'dangerous' level of intelligence, where unaligned operation at a dangerous level of intelligence kills everybody on Earth and then we don't get to try again.**

This includes, for example: (a) something smart enough to build a nanosystem which has been explicitly authorized to build a nanosystem; or (b) something smart enough to build a nanosystem and also smart enough to gain unauthorized access to the Internet and pay a human to put together the ingredients for a nanosystem; or (c) something smart enough to get unauthorized access to the Internet and build something smarter than itself on the number of machines it can hack; or (d) something smart enough to treat humans as manipulable machinery and which has any authorized or unauthorized two-way causal channel with humans; or (e) something smart enough to improve itself enough to do (b) or (d); etcetera. We can gather all sorts of information beforehand *from less powerful systems that will not kill us if we screw up operating them*; but once we are running more powerful systems, we can no longer update on sufficiently catastrophic errors. This is where practically all of the real lethality comes from, that we have to get things right on the first sufficiently-critical try. If we had unlimited retries - if every time an AGI destroyed all the galaxies we got to go back in time four years and try again - we would in a hundred years figure out which bright ideas actually worked. Human beings can figure out pretty difficult things over time, when they get lots of tries; when a failed

guess kills literally everyone, that is harder. That we have to get a bunch of key stuff right *on the first try* is where most of the lethality really and ultimately comes from; likewise the fact that no authority is here to tell us a list of what exactly is 'key' and will kill us if we get it wrong. (One remarks that most people are so absolutely and flatly unprepared by their 'scientific' educations to challenge pre-paradigmatic puzzles with no scholarly authoritative supervision, that they do not even realize how much harder that is, or how incredibly lethal it is to demand getting that right on the first critical try.)

**4. We can't just "decide not to build AGI"** because GPUs are everywhere, and knowledge of algorithms is constantly being improved and published; 2 years after the leading actor has the capability to destroy the world, 5 other actors will have the capability to destroy the world. **The given lethal challenge is to solve within a time limit**, driven by the dynamic in which, over time, increasingly weak actors with a smaller and smaller fraction of total computing power, become able to build AGI and destroy the world. Powerful actors all refraining in unison from doing the suicidal thing just delays this time limit - it does not lift it, unless computer hardware and computer software progress are both brought to complete severe halts across the whole Earth. The current state of this cooperation to have every big actor refrain from doing the stupid thing, is that at present some large actors with a lot of researchers and computing power are led by people who vocally disdain all talk of AGI safety (eg Facebook AI Research). Note that needing to solve AGI alignment *only* within a time limit, but with unlimited safe retries for rapid experimentation on the full-powered system; or *only* on the first critical try, but with an unlimited time bound; would both be terrifically humanity-threatening challenges by historical standards *individually*.

**5. We can't just build a very weak system**, which is less dangerous because it is so weak, and declare victory; because later there will be more actors that have the capability to build a stronger system and one of them will do so. I've also in the past called this the 'safe-but-useless' tradeoff, or 'safe-vs-useful'. People keep on going "why don't we only use AIs to do X, that seems safe" and the answer is almost always either "doing X in fact takes very powerful cognition that is not passively safe" or, even more commonly, "because restricting yourself to doing X will not prevent Facebook AI Research from destroying the world six months later". If all you need is an object that doesn't do dangerous things, you could try a sponge; a sponge is very passively safe. Building a sponge, however, does not prevent Facebook AI Research from destroying the world six months later when they catch up to the leading actor.

**6. We need to align the performance of some large task, a 'pivotal act' that prevents other people from building an unaligned AGI that destroys the world.** While the number of actors with AGI is few or one, they must execute some "pivotal act", strong enough to flip the gameboard, using an AGI powerful enough to do that. It's not enough to be able to align a *weak system* - we need to align a system that can do some single *very large thing*. The example I usually give is "burn all GPUs". This is not what I think you'd actually want to do with a powerful AGI - the nanomachines would need to operate in an incredibly complicated open environment to hunt down all the GPUs, and that would be needlessly difficult to align. However, all known pivotal acts are currently outside the Overton Window, and I expect them to stay there. So I picked an example where if anybody says "how dare you propose burning all GPUs?" I can say "Oh, well, I don't *actually* advocate doing that; it's just a mild overestimate for the rough power level of what you'd have to do, and the rough level of machine cognition required to do that, in order to prevent somebody else from destroying the world in six months or three years." (If it wasn't a mild overestimate, then 'burn all GPUs' would actually be the minimal pivotal task and hence correct)

answer, and I wouldn't be able to give that denial.) Many clever-sounding proposals for alignment fall apart as soon as you ask "How could you use this to align a system that you could use to shut down all the GPUs in the world?" because it's then clear that the system can't do something that powerful, or, if it can do that, the system wouldn't be easy to align. A GPU-burner is also a system powerful enough to, and purportedly authorized to, build nanotechnology, so it requires operating in a dangerous domain at a dangerous level of intelligence and capability; and this goes along with any non-fantasy attempt to name a way an AGI could change the world such that a half-dozen other would-be AGI-builders won't destroy the world 6 months later.

**7.** The reason why nobody in this community has successfully named a 'pivotal weak act' where you do something weak enough with an AGI to be passively safe, but powerful enough to prevent any other AGI from destroying the world a year later - and yet also we can't just go do that right now and need to wait on AI - is that *nothing like that exists*. There's no reason why it should exist. There is not some elaborate clever reason why it exists but nobody can see it. It takes a lot of power to do something to the current world that prevents any other AGI from coming into existence; nothing which can do that is passively safe in virtue of its weakness. If you can't solve the problem right now (which you can't, because you're opposed to other actors who don't want to be solved and those actors are on roughly the same level as you) then you are resorting to some cognitive system that can do things you could not figure out how to do yourself, that you were not *close* to figuring out because you are not *close* to being able to, for example, burn all GPUs. Burning all GPUs would *actually* stop Facebook AI Research from destroying the world six months later; weaksauce Overton-abiding stuff about 'improving public epistemology by setting GPT-4 loose on Twitter to provide scientifically literate arguments about everything' will be cool but will not actually prevent Facebook AI Research from destroying the world six months later, or some eager open-source collaborative from destroying the world a year later if you manage to stop FAIR specifically. **There are no pivotal weak acts.**

**8. The best and easiest-found-by-optimization algorithms for solving problems we want an AI to solve, readily generalize to problems we'd rather the AI not solve;** you can't build a system that only has the capability to drive red cars and not blue cars, because all red-car-driving algorithms generalize to the capability to drive blue cars.

**9.** The builders of a safe system, by hypothesis on such a thing being possible, would need to operate their system in a regime where it has the *capability* to kill everybody or make itself even more dangerous, but has been successfully designed to not do that. **Running AGIs doing something pivotal are not passively safe**, they're the equivalent of nuclear cores that require actively maintained design properties to not go supercritical and melt down.

## Section B:

Okay, but as we all know, modern machine learning is like a genie where you just give it a wish, right? Expressed as some mysterious thing called a 'loss function', but which is basically just equivalent to an English wish phrasing, right? And then if you pour in enough computing power you get your wish, right? So why not train a giant stack of transformer layers on a dataset of agents doing nice things and not bad

things, throw in the word 'corrigibility' somewhere, crank up that computing power, and get out an aligned AGI?

### Section B.1: The distributional leap.

**10.** You can't train alignment by running lethally dangerous cognitions, observing whether the outputs kill or deceive or corrupt the operators, assigning a loss, and doing supervised learning. **On anything like the standard ML paradigm, you would need to somehow generalize optimization-for-alignment you did in safe conditions, across a big distributional shift to dangerous conditions.** (Some generalization of this seems like it would have to be true even outside that paradigm; you wouldn't be working on a live unaligned superintelligence to align it.) This alone is a point that is sufficient to kill a lot of naive proposals from people who never did or could concretely sketch out any specific scenario of what training they'd do, in order to align what output - which is why, of course, they never concretely sketch anything like that. **Powerful AGIs doing dangerous things that will kill you if misaligned, must have an alignment property that generalized far out-of-distribution from safer building/training operations that didn't kill you.**

This is where a huge amount of lethality comes from on anything remotely resembling the present paradigm. Unaligned operation at a dangerous level of intelligence\*capability will kill you; so, if you're starting with an unaligned system and labeling outputs in order to get it to learn alignment, the training regime or building regime must be operating at some lower level of intelligence\*capability that is passively safe, where its currently-unaligned operation does not pose any threat. (Note that anything substantially smarter than you poses a threat given *any* realistic level of capability. Eg, "being able to produce outputs that humans look at" is probably sufficient for a generally much-smarter-than-human AGI to [navigate its way out of the causal systems that are humans](#), especially in the real world where somebody trained the system on terabytes of Internet text, rather than somehow keeping it ignorant of the latent causes of its source code and training environments.)

**11.** If cognitive machinery doesn't generalize far out of the distribution where you did tons of training, it can't solve problems on the order of 'build nanotechnology' where it would be too expensive to run a million training runs of failing to build nanotechnology. There is no pivotal act this weak; **there's no known case where you can entrain a safe level of ability on a safe environment where you can cheaply do millions of runs, and deploy that capability to save the world** and prevent the next AGI project up from destroying the world two years later. Pivotal weak acts like this aren't known, and not for want of people looking for them. So, again, you end up needing alignment to generalize way out of the training distribution - not just because the training environment needs to be safe, but because the training environment probably also needs to be *cheaper* than evaluating some real-world domain in which the AGI needs to do some huge act. You don't get 1000 failed tries at burning all GPUs - because people will notice, even leaving out the consequences of capabilities success and alignment failure.

**12. Operating at a highly intelligent level is a drastic shift in distribution from operating at a less intelligent level,** opening up new external options, and probably opening up even more new internal choices and modes. Problems that materialize at high intelligence and danger levels may fail to show up at safe lower levels of intelligence, or may recur after being suppressed by a first patch.

**13. Many alignment problems of superintelligence will not naturally appear at pre-dangerous, passively-safe levels of capability.** Consider the internal behavior 'change your outer behavior to deliberately look more aligned and deceive the programmers, operators, and possibly any loss functions optimizing over you'. This problem is one that will appear at the superintelligent level; if, being otherwise ignorant, we guess that it is among the *median* such problems in terms of how *early* it naturally appears in earlier systems, then around *half* of the alignment problems of superintelligence will first naturally materialize *after* that one first starts to appear. Given *correct* foresight of which problems will naturally materialize *later*, one could try to deliberately materialize such problems earlier, and get in some observations of them. This helps to the extent (a) that we actually correctly forecast all of the problems that will appear later, or some superset of those; (b) that we succeed in preemptively materializing a superset of problems that will appear later; and (c) that we can actually solve, in the earlier laboratory that is out-of-distribution for us relative to the real problems, those alignment problems that would be lethal if we mishandle them when they materialize later. Anticipating *all* of the really dangerous ones, and then successfully materializing them, in the correct form for early solutions to generalize over to later solutions, *sounds possibly kinda hard*.

**14. Some problems**, like 'the AGI has an option that (looks to it like) it could successfully kill and replace the programmers to fully optimize over its environment', **seem like their natural order of appearance could be that they first appear only in fully dangerous domains**. Really actually having a *clear* option to brain-level-persuade the operators or escape onto the Internet, build nanotech, and destroy all of humanity - in a way where you're fully clear that you know the relevant facts, and estimate only a not-worth-it low probability of learning something which changes your preferred strategy if you bide your time another month while further growing in capability - is an option that first gets evaluated for real at the point where an AGI fully expects it can defeat its creators. We can try to manifest an echo of that apparent scenario in earlier toy domains. Trying to train by gradient descent against that behavior, in that toy domain, is something I'd expect to produce not-particularly-coherent local patches to thought processes, which would break with near-certainty inside a superintelligence generalizing far outside the training distribution and thinking very different thoughts. Also, programmers and operators themselves, who are used to operating in not-fully-dangerous domains, are operating out-of-distribution when they enter into dangerous ones; our methodologies may at that time break.

**15. Fast capability gains seem likely, and may break lots of previous alignment-required invariants simultaneously.** Given otherwise insufficient foresight by the operators, I'd expect a lot of those problems to appear approximately simultaneously after a sharp capability gain. See, again, the case of human intelligence. We didn't break alignment with the 'inclusive reproductive fitness' outer loss function, immediately after the introduction of farming - something like 40,000 years into a 50,000 year Cro-Magnon takeoff, as was itself running very quickly relative to the outer optimization loop of natural selection. Instead, we got a lot of technology more advanced than was in the ancestral environment, including contraception, in one very fast burst relative to the speed of the outer optimization loop, late in the general intelligence game. We started reflecting on ourselves a lot more, started being programmed a lot more by cultural evolution, and lots and lots of assumptions underlying our alignment in the ancestral training environment broke simultaneously. (People will perhaps rationalize reasons why this abstract description doesn't carry over to gradient descent; eg, "gradient descent has less of an information bottleneck". My model of this variety of reader has an inside view, which

they will label an outside view, that assigns great relevance to some other data points that are *not* observed cases of an outer optimization loop producing an inner general intelligence, and assigns little importance to our one data point actually featuring the phenomenon in question. When an outer optimization loop actually produced general intelligence, it broke alignment after it turned general, and did so relatively late in the game of that general intelligence accumulating capability and knowledge, almost immediately before it turned 'lethally' dangerous relative to the outer optimization loop of natural selection. Consider skepticism, if someone is ignoring this one warning, especially if they are not presenting equally lethal and dangerous things that they say will go wrong instead.)

### **Section B.2: Central difficulties of outer and inner alignment.**

**16.** Even if you train really hard on an exact loss function, that doesn't thereby create an explicit internal representation of the loss function inside an AI that then continues to pursue that exact loss function in distribution-shifted environments. Humans don't explicitly pursue inclusive genetic fitness; **outer optimization even on a very exact, very simple loss function doesn't produce inner optimization in that direction.** This happens *in practice in real life*, it is what happened in *the only case we know about*, and it seems to me that there are deep theoretical reasons to expect it to happen again: the *first* semi-outer-aligned solutions found, in the search ordering of a real-world bounded optimization process, are not inner-aligned solutions. This is sufficient on its own, even ignoring many other items on this list, to trash entire categories of naive alignment proposals which assume that if you optimize a bunch on a loss function calculated using some simple concept, you get perfect inner alignment on that concept.

**17.** More generally, a superproblem of 'outer optimization doesn't produce inner alignment' is that **on the current optimization paradigm there is no general idea of how to get particular inner properties into a system, or verify that they're there, rather than just observable outer ones you can run a loss function over.** This is a problem when you're trying to generalize out of the original training distribution, because, eg, the outer behaviors you see could have been produced by an inner-misaligned system that is deliberately producing outer behaviors that will fool you. We don't know how to get any bits of information into the *inner* system rather than the *outer* behaviors, in any systematic or general way, on the current optimization paradigm.

**18. There's no reliable Cartesian-sensory ground truth** (reliable loss-function-calculator) **about whether an output is 'aligned'**, because some outputs destroy (or fool) the human operators and produce a different environmental causal chain behind the externally-registered loss function. That is, if you show an agent a reward signal that's currently being generated by humans, the signal is not *in general a reliable perfect ground truth about how aligned an action was*, because another way of producing a high reward signal is to deceive, corrupt, or replace the human operators with a different causal system which generates that reward signal. When you show an agent an environmental reward signal, you are not showing it something that is a reliable ground truth about whether the system did the thing you wanted it to do; *even if* it ends up perfectly inner-aligned on that reward signal, or learning some concept that *exactly* corresponds to 'wanting states of the environment which result in a high reward signal being sent', an AGI strongly optimizing on that signal will kill you,

because the sensory reward signal was not a ground truth about alignment (as seen by the operators).

**19.** More generally, **there is no known way to use the paradigm of loss functions, sensory inputs, and/or reward inputs, to optimize anything within a cognitive system to point at particular things within the environment** - to point to *latent events and objects and properties in the environment*, rather than *relatively shallow functions of the sense data and reward*. This isn't to say that nothing in the system's goal (whatever goal accidentally ends up being inner-optimized over) could ever point to anything in the environment by *accident*. Humans ended up pointing to their environments at least partially, though we've got lots of internally oriented motivational pointers as well. But insofar as the current paradigm works at all, the on-paper design properties say that it only works for aligning on known direct functions of sense data and reward functions. All of these kill you if optimized-over by a sufficiently powerful intelligence, because they imply strategies like 'kill everyone in the world using nanotech to strike before they know they're in a battle, and have control of your reward button forever after'. It just isn't *true* that we know a function on webcam input such that every world with that webcam showing the right things is safe for us creatures outside the webcam. This general problem is a fact about the territory, not the map; it's a fact about the actual environment, not the particular optimizer, that lethal-to-us possibilities exist in some possible environments underlying every given sense input.

**20.** Human operators are fallible, breakable, and manipulable. **Human raters make systematic errors - regular, compactly describable, predictable errors.**

To *faithfully* learn a function from 'human feedback' is to learn (from our external standpoint) an unfaithful description of human preferences, with errors that are not random (from the outside standpoint of what we'd hoped to transfer). If you perfectly learn and perfectly maximize *the referent* of rewards assigned by human operators, that kills them. It's a fact about the territory, not the map - about the environment, not the optimizer - that the *best predictive* explanation for human answers is one that predicts the systematic errors in our responses, and therefore is a psychological concept that correctly predicts the higher scores that would be assigned to human-error-producing cases.

**21.** There's something like a single answer, or a single bucket of answers, for questions like 'What's the environment really like?' and 'How do I figure out the environment?' and 'Which of my possible outputs interact with reality in a way that causes reality to have certain properties?', where a simple outer optimization loop will straightforwardly shove optimizers into this bucket. When you have a wrong belief, reality hits back at your wrong predictions. When you have a broken belief-updater, reality hits back at your broken predictive mechanism via predictive losses, and a gradient descent update fixes the problem in a simple way that can easily cohere with all the other predictive stuff. In contrast, when it comes to a choice of utility function, there are unbounded degrees of freedom and multiple reflectively coherent fixpoints. Reality doesn't 'hit back' against things that are locally aligned with the loss function on a particular range of test cases, but globally misaligned on a wider range of test cases. This is the very abstract story about why hominids, once they finally started to generalize, generalized their *capabilities* to Moon landings, but their inner optimization no longer adhered very well to the outer-optimization goal of 'relative inclusive reproductive fitness' - even though they were in their ancestral environment optimized very strictly around this one thing and nothing else. This abstract dynamic is something you'd expect to be true about outer optimization loops on the order of

both 'natural selection' and 'gradient descent'. The central result: **Capabilities generalize further than alignment once capabilities start to generalize far.**

**22.** There's a relatively simple core structure that explains why complicated cognitive machines work; which is why such a thing as general intelligence exists and not just a lot of unrelated special-purpose solutions; which is why capabilities generalize after outer optimization infuses them into something that has been optimized enough to become a powerful inner optimizer. The fact that this core structure is simple and relates generically to [low-entropy high-structure environments](#) is why humans can walk on the Moon. **There is no analogous truth about there being a simple core of alignment**, especially not one that is even easier for gradient descent to find than it would have been for natural selection to just find 'want inclusive reproductive fitness' as a well-generalizing solution within ancestral humans. Therefore, capabilities generalize further out-of-distribution than alignment, once they start to generalize at all.

**23. Corrigibility is anti-natural to consequentialist reasoning;** "you can't bring the coffee if you're dead" for almost every kind of coffee. We (MIRI) [tried and failed](#) to find a coherent formula for an agent that would let itself be shut down (without that agent actively trying to get shut down). Furthermore, many anti-corrigible lines of reasoning like this may only first appear at high levels of intelligence.

**24.** There are two fundamentally different approaches you can potentially take to alignment, which are unsolvable for two different sets of reasons; therefore, **by becoming confused and ambiguating between the two approaches, you can confuse yourself about whether alignment is necessarily difficult.** The first approach is to build a CEV-style Sovereign which wants exactly what we extrapolated-want and is therefore safe to let optimize all the future galaxies without it accepting any human input trying to stop it. The second course is to build corrigible AGI which doesn't want exactly what we want, and yet somehow fails to kill us and take over the galaxies despite that being a convergent incentive there.

1. The first thing generally, or CEV specifically, is unworkable because **the complexity of what needs to be aligned or meta-aligned for our Real Actual Values is far out of reach for our FIRST TRY at AGI**. Yes I mean specifically that the *dataset, meta-learning algorithm, and what needs to be learned*, is far out of reach for our first try. It's not just non-hand-codable, it is *unteachable* on-the-first-try because *the thing you are trying to teach is too weird and complicated*.
2. The second thing looks unworkable (less so than CEV, but still lethally unworkable) because **corrigibility runs actively counter to instrumentally convergent behaviors** within a core of general intelligence (the capability that generalizes far out of its original distribution). You're not trying to make it have an opinion on something the core was previously neutral on. You're trying to take a system implicitly trained on lots of arithmetic problems until its machinery started to reflect the common coherent core of arithmetic, and get it to say that as a special case  $222 + 222 = 555$ . You can maybe train something to do this in a particular training distribution, but it's incredibly likely to break when you present it with new math problems far outside that training distribution, on a system which successfully generalizes capabilities that far at all.

### **Section B.3: Central difficulties of *sufficiently good and useful* transparency / interpretability.**

**25. We've got no idea what's actually going on inside the giant inscrutable matrices and tensors of floating-point numbers.** Drawing interesting graphs of where a transformer layer is focusing attention doesn't help if the question that needs answering is "So was it planning how to kill us or not?"

**26.** Even if we did know what was going on inside the giant inscrutable matrices while the AGI was still too weak to kill us, this would just result in us dying with more dignity, if DeepMind refused to run that system and let Facebook AI Research destroy the world two years later. **Knowing that a medium-strength system of inscrutable matrices is planning to kill us, does not thereby let us build a high-strength system of inscrutable matrices that isn't planning to kill us.**

**27.** When you explicitly optimize against a detector of unaligned thoughts, you're partially optimizing for more aligned thoughts, and partially optimizing for unaligned thoughts that are harder to detect. **Optimizing against an interpreted thought optimizes against interpretability.**

**28.** The AGI is smarter than us in whatever domain we're trying to operate it inside, so we cannot mentally check all the possibilities it examines, and we cannot see all the consequences of its outputs using our own mental talent. **A powerful AI searches parts of the option space we don't, and we can't foresee all its options.**

**29.** The outputs of an AGI go through a huge, not-fully-known-to-us domain (the real world) before they have their real consequences. **Human beings cannot inspect an AGI's output to determine whether the consequences will be good.**

**30.** Any pivotal act that is not something we can go do right now, will take advantage of the AGI figuring out things about the world we don't know so that it can make plans we wouldn't be able to make ourselves. It knows, at the least, the fact we didn't previously know, that some action sequence results in the world we want. Then humans will not be competent to use their own knowledge of the world to figure out all the results of that action sequence. An AI whose action sequence you can fully understand all the effects of, before it executes, is much weaker than humans in that domain; you couldn't make the same guarantee about an unaligned human as smart as yourself and trying to fool you. **There is no pivotal output of an AGI that is humanly checkable and can be used to safely save the world but only after checking it;** this is another form of pivotal weak act which does not exist.

**31.** A strategically aware intelligence can choose its visible outputs to have the consequence of deceiving you, including about such matters as whether the intelligence has acquired strategic awareness; **you can't rely on behavioral inspection to determine facts about an AI which that AI might want to deceive you about.** (Including how smart it is, or whether it's acquired strategic awareness.)

**32.** Human thought partially exposes only a partially scrutable outer surface layer. Words only trace our real thoughts. Words are not an AGI-complete data representation in its native style. The underparts of human thought are not exposed for direct imitation learning and can't be put in any dataset. **This makes it hard and probably impossible to train a powerful system entirely on imitation of human words or other human-legible contents,** which are only impoverished

subsystems of human thoughts; **unless that system is powerful enough to contain inner intelligences figuring out the humans**, and at that point it is no longer really working as imitative human thought.

**33. The AI does not think like you do**, the AI doesn't have thoughts built up from the same concepts you use, it is utterly alien on a staggering scale. Nobody knows what the hell GPT-3 is thinking, not *only* because the matrices are opaque, but because the *stuff within that opaque container* is, very likely, incredibly alien - nothing that would translate well into comprehensible human thinking, even if we could see past the giant wall of floating-point numbers to what lay behind.

#### **Section B.4: Miscellaneous unworkable schemes.**

**34. Coordination schemes between superintelligences are not things that humans can participate in** (eg because humans can't reason reliably about the code of superintelligences); a "multipolar" system of 20 superintelligences with different utility functions, plus humanity, has a natural and obvious equilibrium which looks like "the 20 superintelligences cooperate with each other but not with humanity".

**35.** Schemes for playing "different" AIs off against each other stop working if those AIs advance to the point of being able to coordinate via reasoning about (probability distributions over) each others' code. **Any system of sufficiently intelligent agents can probably behave as a single agent, even if you imagine you're playing them against each other.** Eg, if you set an AGI that is secretly a paperclip maximizer, to check the output of a nanosystems designer that is secretly a staples maximizer, then even if the nanosystems designer is not able to deduce what the paperclip maximizer really wants (namely paperclips), it could still logically commit to share half the universe with any agent checking its designs if those designs were allowed through, *if* the checker-agent can verify the suggester-system's logical commitment and hence logically depend on it (which excludes human-level intelligences). Or, if you prefer simplified catastrophes without any logical decision theory, the suggester could bury in its nanosystem design the code for a new superintelligence that will visibly (to a superhuman checker) divide the universe between the nanosystem designer and the design-checker.

**36.** What makes an air conditioner 'magic' from the perspective of say the thirteenth century, is that even if you correctly show them the design of the air conditioner in advance, they won't be able to understand from seeing that design why the air comes out cold; the design is exploiting regularities of the environment, rules of the world, laws of physics, that they don't know about. The domain of human thought and human brains is very poorly understood by us, and exhibits phenomena like optical illusions, hypnosis, psychosis, mania, or simple afterimages produced by strong stimuli in one place leaving neural effects in another place. Maybe a superintelligence couldn't defeat a human in a very simple realm like logical tic-tac-toe; if you're fighting it in an incredibly complicated domain you understand poorly, like human minds, you should expect to be defeated by 'magic' in the sense that even if you saw its strategy you would not understand why that strategy worked. **AI-boxing can only work on relatively weak AGIs; the human operators are not secure systems.**

## Section C:

Okay, those are some significant problems, but lots of progress is being made on solving them, right? There's a whole field calling itself "AI Safety" and many major organizations are expressing Very Grave Concern about how "safe" and "ethical" they are?

**37.** There's a pattern that's played out quite often, over all the times the Earth has spun around the Sun, in which some bright-eyed young scientist, young engineer, young entrepreneur, proceeds in full bright-eyed optimism to challenge some problem that turns out to be really quite difficult. Very often the cynical old veterans of the field try to warn them about this, and the bright-eyed youngsters don't listen, because, like, who wants to hear about all that stuff, they want to go solve the problem! Then this person gets beaten about the head with a slipper by reality as they find out that their brilliant speculative theory is wrong, it's actually really hard to build the thing because it keeps breaking, and society isn't as eager to adopt their clever innovation as they might've hoped, in a process which eventually produces a new cynical old veteran. Which, if not literally optimal, is I suppose a nice life cycle to nod along to in a nature-show sort of way. Sometimes you do something for the *first* time and there are no cynical old veterans to warn anyone and people can be *really* optimistic about how it will go; eg the initial Dartmouth Summer Research Project on Artificial Intelligence in 1956: "An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer." This is *less* of a viable survival plan for your *planet* if the first major failure of the bright-eyed youngsters kills *literally everyone* before they can predictably get beaten about the head with the news that there were all sorts of unforeseen difficulties and reasons why things were hard. You don't get any cynical old veterans, in this case, because everybody on Earth is dead. Once you start to suspect you're in that situation, you have to do the Bayesian thing and update now to the view you will predictably update to later: realize you're in a situation of being that bright-eyed person who is going to encounter Unexpected Difficulties later and end up a cynical old veteran - or would be, except for the part where you'll be dead along with everyone else. And become that cynical old veteran *right away*, before reality whaps you upside the head in the form of everybody dying and you not getting to learn. **Everyone else seems to feel that, so long as reality hasn't whapped them upside the head yet and smacked them down with the actual difficulties, they're free to go on living out the standard life-cycle and play out their role in the script and go on being bright-eyed youngsters; there's no cynical old veterans to warn them otherwise, after all, and there's no proof that everything won't go beautifully easy and fine, given their bright-eyed total ignorance of what those later difficulties could be.**

**38. It does not appear to me that the field of 'AI safety' is currently being remotely productive on tackling its enormous lethal problems.** These problems are in fact out of reach; the contemporary field of AI safety has been selected to contain people who go to work in that field anyways. Almost all of them are there to tackle problems on which they can appear to succeed and publish a paper claiming success; if they can do that and get funded, why would they embark on a much more unpleasant project of trying something harder that they'll fail at, just so

the human species can die with marginally more dignity? This field is not making real progress and does not have a recognition function to distinguish real progress if it took place. You could pump a billion dollars into it and it would produce mostly noise to drown out what little progress was being made elsewhere.

**39. I figured this stuff out using the [null string](#) as input**, and frankly, I have a hard time myself feeling hopeful about getting real alignment work out of somebody who previously sat around waiting for somebody else to input a persuasive argument into them. This ability to "notice lethal difficulties without Eliezer Yudkowsky arguing you into noticing them" currently is an opaque piece of cognitive machinery to me, I do not know how to train it into others. It probably relates to '[security mindset](#)', and a mental motion where you refuse to play out scripts, and being able to operate in a field that's in a state of chaos.

**40.** "Geniuses" with nice legible accomplishments in fields with tight feedback loops where it's easy to determine which results are good or bad right away, and so validate that this person is a genius, are (a) people who might not be able to do equally great work away from tight feedback loops, (b) people who chose a field where their genius would be nicely legible even if that maybe wasn't the place where humanity most needed a genius, and (c) probably don't have the mysterious gears simply because they're rare. **You cannot just pay \$5 million apiece to a bunch of legible geniuses from other fields and expect to get great alignment work out of them.** They probably do not know where the real difficulties are, they probably do not understand what needs to be done, *they cannot tell the difference between good and bad work*, and the funders also can't tell without me standing over their shoulders evaluating everything, which I do not have the physical stamina to do. I concede that real high-powered talents, especially if they're still in their 20s, genuinely interested, and have done their reading, are people who, yeah, fine, have higher probabilities of making core contributions than a random bloke off the street. But I'd have more hope - not significant hope, but *more* hope - in separating the concerns of (a) credibly promising to pay big money retrospectively for good work to anyone who produces it, and (b) venturing prospective payments to somebody who is predicted to maybe produce good work later.

**41. Reading this document cannot make somebody a core alignment researcher.** That requires, not the ability to read this document and nod along with it, but the ability to spontaneously write it from scratch without anybody else prompting you; that is what makes somebody a peer of its author. It's guaranteed that some of my analysis is mistaken, though not necessarily in a hopeful direction. The ability to do new basic work noticing and fixing those flaws is the same ability as the ability to write this document before I published it, which nobody apparently did, despite my having had other things to do than write this up for the last five years or so. Some of that silence may, possibly, optimistically, be due to nobody else in this field having the ability to write things comprehensibly - such that somebody out there had the knowledge to write all of this themselves, if they could only have written it up, but they couldn't write, so didn't try. I'm not particularly hopeful of this turning out to be true in real life, but I suppose it's one possible place for a "positive model violation" (miracle). The fact that, twenty-one years into my entering this death game, seven years into other EAs noticing the death game, and two years into even normies starting to notice the death game, it is still Eliezer Yudkowsky writing up this list, says that humanity still has only one gamepiece that can do that. I knew I did not actually have the physical stamina to be a star researcher, I tried really really hard to replace myself before my health deteriorated further, and yet here I am writing this. That's not what surviving worlds look like.

**42. There's no plan.** Surviving worlds, by this point, and in fact several decades earlier, have a plan for how to survive. It is a written plan. The plan is not secret. In this non-surviving world, there are no candidate plans that do not immediately fall to Eliezer instantly pointing at the giant visible gaping holes in that plan. Or if you don't know who Eliezer is, you don't even realize you need a plan, because, like, how would a human being possibly realize that without Eliezer yelling at them? It's not like people will yell at *themselves* about prospective alignment difficulties, they don't have an *internal* voice of caution. So most organizations don't have plans, because I haven't taken the time to personally yell at them. 'Maybe we should have a plan' is deeper alignment mindset than they possess without me standing constantly on their shoulder as their personal angel pleading them into... continued noncompliance, in fact. Relatively few are aware even that they should, to look better, produce a *pretend* plan that can fool EAs too '[modest](#)' to trust their own judgments about seemingly gaping holes in what serious-looking people apparently believe.

**43. This situation you see when you look around you is not what a surviving world looks like.** The worlds of humanity that survive have plans. They are not leaving to one tired guy with health problems the entire responsibility of pointing out real and lethal problems proactively. Key people are taking internal and real responsibility for finding flaws in their own plans, instead of considering it their job to propose solutions and somebody else's job to prove those solutions wrong. That world started trying to solve their important lethal problems earlier than this. Half the people going into string theory shifted into AI alignment instead and made real progress there. When people suggest a planetarily-lethal problem that might materialize later - there's a lot of people suggesting those, in the worlds destined to live, and they don't have a special status in the field, it's just what normal geniuses there do - they're met with either solution plans or a reason why that shouldn't happen, not an uncomfortable shrug and 'How can you be sure that will happen' / 'There's no way you could be sure of that now, we'll have to wait on experimental evidence.'

A lot of those better worlds will die anyways. It's a genuinely difficult problem, to solve something like that on your first try. But they'll die with more dignity than this.

# It's Probably Not Lithium

*This post has been recorded as part of the LessWrong Curated Podcast, and can be listened to on [Spotify](#), [Apple Podcasts](#), [Libsyn](#), and more.*

---

[A Chemical Hunger](#) (a), a series by the authors of the blog [Slime Mold Time Mold](#) (SMTM) that [has been received positively on LessWrong](#), argues that the obesity epidemic is [entirely caused](#) (a) by environmental contaminants. The authors' top suspect [is lithium](#) (a)<sup>[1]</sup>, primarily because it is known to cause weight gain at the doses used to treat bipolar disorder.

After doing some research, however, I found that it is not plausible that lithium plays a major role in the obesity epidemic, and that a lot of the claims the SMTM authors make about the topic are misleading, flat-out wrong, or based on extremely cherry-picked evidence. I have the impression that reading what they have to say about this often leaves the reader with a worse model of reality than they started with, and I'll explain why I have that impression in this post.

## (Preamble) A brief summary of their hypotheses

The SMTM authors have [recently](#) (a) summarized their hypotheses on how lithium exposure could explain the obesity epidemic. The first hypothesis is that trace exposure is responsible:

One possibility is that small amounts of lithium are enough to cause obesity, at least with daily exposure.

And the second one is that people are intermittently exposed to therapeutic doses:

[E]ven if people aren't getting that much lithium **on average**, if they **sometimes** get huge doses, that could be enough to drive their lipostat upward.

I am going to argue that **neither of those is plausible**. I address the plausibility of the second hypothesis in the next section, and the plausibility of the first one in the rest of the post.

## Lithium exposure in the general population is extremely low, even at the tails, in the majority of countries for which we have data

A few days ago, the SMTM authors published [a literature review](#) (a) on the lithium content of food. They conclude that, whereas the existing literature isn't great, "[i]t seems like most people get at least 1 mg [of lithium] a day from their food, and on

many days, there's a good chance you'll get more." They also say it seems plausible that people are intermittently exposed to doses of lithium within the therapeutic range through their diet.

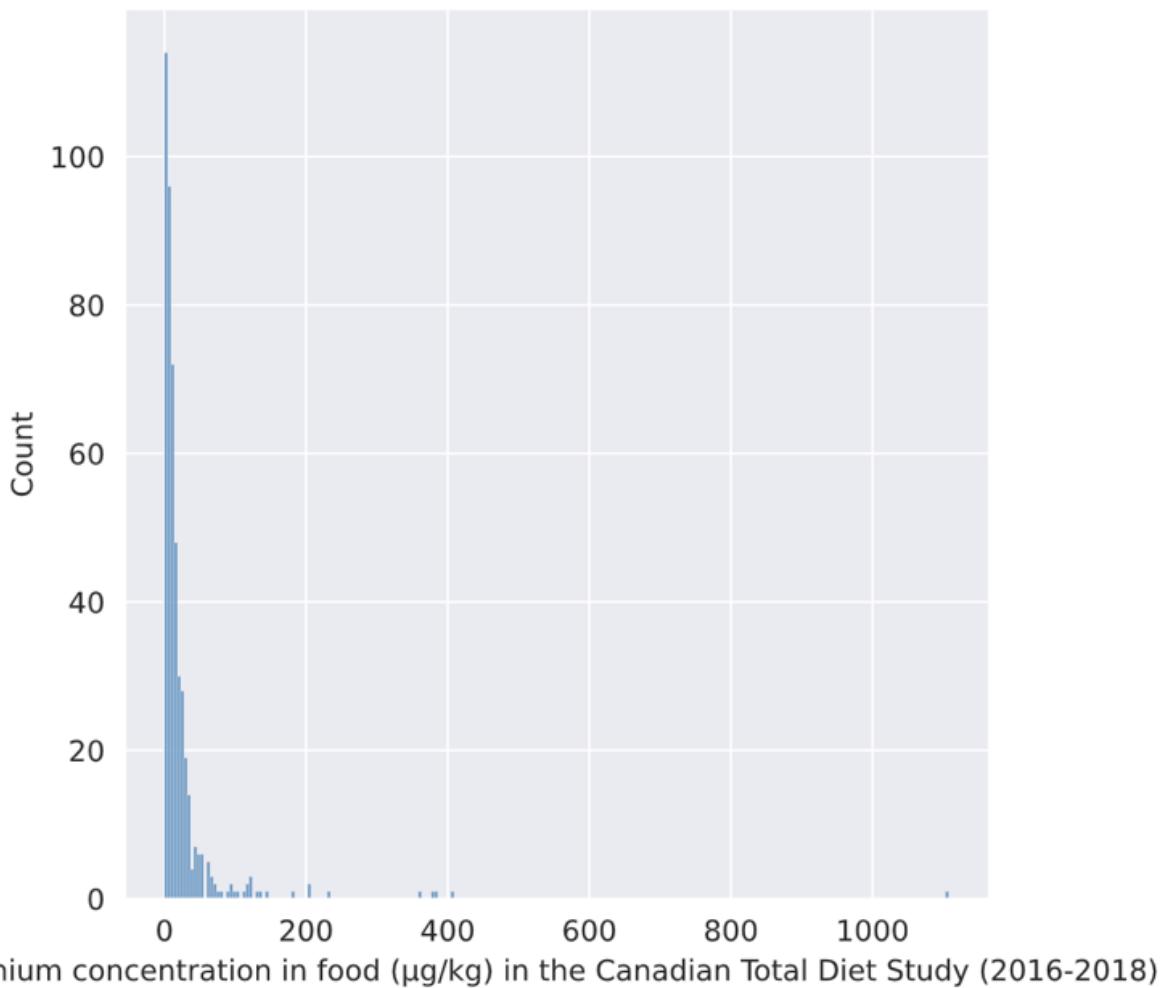
However, their literature review *pretty much only includes studies that are outliers in the literature*. Moreover, they use a misleading threshold for the therapeutic range of lithium. I'll explain.

## **The studies in SMTM's literature review of lithium levels in food are pretty much all outliers**

In 2006, France conducted its second Total Diet Study (henceforth TDS). Across 1,319 food samples, the highest lithium concentration found was 0.6 mg/kg, in water. That's not the highest average concentration among food groups – it's the highest concentration of any single sample they tested. (For context, a standard clinical dose of elemental lithium is about 200 mg/day, or 1 gram/day of lithium carbonate.)

Similarly, New Zealand's 2016 TDS examined 1,056 food samples and the highest concentration it found in any single sample was 0.54 mg/kg (in mussels).

Canada makes the raw data of its Total Diet Study publicly available (a), and they too measure the lithium content of their food. The maximum level reported is 1.1 mg/kg (in table salt, which is presumably rarely consumed in kilogram quantities) across 479 food samples, with the mean being 25 µg/kg and the median 11 µg/kg. Here's a histogram of the data:



Excluding table salt, the maximum value in the rest of the dataset ( $N = 476$ ) is 0.4  $\mu\text{g}/\text{kg}$ , in mineral water.

Total Diet Studies in other countries report similarly low levels. Using data from the UK's 1994 TDS (which included 400 food samples), the mean daily lithium intake among adults was [estimated](#) to be 17  $\mu\text{g}/\text{day}$ , more than 50 times lower than SMTM's estimate of "at least 1 mg a day". In Northern Italy, a research group that analyzed 908 samples of food from 2016 to 2017 [has estimated](#) that the mean dietary intake is 18.15  $\mu\text{g}/\text{day}$ , and an [older study](#) in the same area, with 248 food samples, estimated daily dietary intake to be 29.9  $\mu\text{g}$ . Those numbers are similar to estimates from France, which range from [11  \$\mu\text{g}/\text{day}\$](#)  to [48  \$\mu\text{g}/\text{day}\$](#) , and to estimates from [New Zealand's 2016 TDS](#), the *upper bound* of which is 0.31  $\mu\text{g}$  per kilogram of body weight per day, or 31  $\mu\text{g}/\text{day}$  for a 220 lb adult.

(Notably, dietary lithium intake in Vietnam, which is extremely lean compared to all other countries mentioned here, [is estimated to be](#) 0.285  $\mu\text{g}$  per kilogram of body weight per day, perfectly in line with these other estimates.)

I haven't been able to find TDS data from other countries. The United States, notably, does not track lithium in its own TDS.<sup>[2]</sup> But there are a few smaller studies available, neatly summarized by [Van Cauwenbergh et al. \(1999\)](#) (including one from the US, in the last row):

**Table 3** Literature data on daily Li intake ( $\mu\text{g}$ )

Country	Sampling technique	Population	Daily intake $x \pm \text{SD}$ (range)	Ref.
Belgium	Duplicate diets	Four different areas (healthy people)	$8.6 \pm 4.6$	This study
Canada	Analysed data; household food survey	Average diets	21.6	18
Germany (DDR)	Duplicate diet	Seven men	(182–546)	1
Italy	Fifteen total diets	Seven women	(220–532)	1
Japan	Duplicate portion, food tables	25–30 years ( $\sigma$ )	$27.8 \pm 8.2$ (16.2–44.6)	19
Spain	Twenty total diets	Non-smoking adults ( $\sigma$ )	<1 <sup>a</sup>	8
Turkey	Six total diets	25–30 years ( $\sigma$ )	$53.5 \pm 25.3$ (10.9–104.7)	19
UK	—	Adults	$39.3 \pm 10.3$ (29.3–50.9)	19
USA	Five total diets	25–30 years ( $\sigma$ )	107	20
			$37.4 \pm 14.6$ (25–62)	19

<sup>a</sup> Median value

*Notice that none of those studies estimates anything as high as “1 mg a day.”*

Older sources not included in that table [suggest](#) that Finnish diets provide 35  $\mu\text{g}/\text{day}$  of lithium, Turkish diets 102  $\mu\text{g}/\text{day}$ , and American diets 60–70  $\mu\text{g}/\text{day}$ .

[SMTM’s review of lithium in food](#) (a) does not include any of these studies; instead, it largely relies on old data from a single author from Germany, a country which, as we can see in the chart above, is a clear outlier. So, unsurprisingly, the numbers that you see in SMTM’s review are way higher than the numbers I found in my own research. The lowest estimate their review even *mentions, as a lower bound*, for daily lithium intake in the general population, is 128  $\mu\text{g}$ .

To be clear, I don’t think that their data is *wrong*. It makes sense for lithium concentration to vary a lot in time and space. [3] And whereas the SMTM authors based their literature review mostly on old German studies, the studies I found were from Canada, France, New Zealand, Italy, the UK, Belgium, etc. and a lot of those are very recent. But it is odd that **they exclusively talk about outliers**, and make conclusions such as “[i]t seems like most people get at least 1 mg a day from their food” *based on those outliers*.

I have attempted to make a comment on SMTM’s post linking to many of those studies, but they have not approved the comment. I have also attempted to contact them on [Twitter](#) (twice) and through email, but have not received a reply. All of this was over one week ago, and they have, since then, replied to other people on Twitter and approved other comments on their post, but haven’t commented on this. So I have no idea why their literature review excludes these studies.

[ETA: on July 5, a week after the publication of this post, they released [a post](#) (a) addressing five of the studies I’ve mentioned here (and completely ignoring the rest). I replied to this new post in [these comments](#).]

(Notably, the SMTM authors say in the post that “the smart money is that Anke’s measurements [Anke being the German author most of their review is based on] are probably all lower than the levels in modern food.” If my data from Canada, New Zealand and Italy, all of which is from after 2015, counts as “modern,” then that prediction seems to have turned out horribly wrong.)

Their literature review was misleading in other ways as well; I explain how in [my addenda](#).

---

People eat about 2 kg of food per day. So it seems that, at least in New Zealand, Canada and France, in order to consume 1.2 mg of lithium in a day you'd need to spend the entire day eating nothing but the most lithium-rich of the 2,851 food samples tested in those three countries combined (excluding salt). And even *that* value is still 100 times lower than a low therapeutic dose (which, [as I'll explain below](#), is about 100 mg/day of elemental lithium). So SMTM's hypothesis that people are intermittently exposed to doses of lithium within the therapeutic range through their diet seems very implausible, at least in those countries (in which, just to make it clear, the obesity epidemic has [definitely arrived \(a\)](#)). <sup>[4]</sup>

---

This isn't, [and could not be](#), conclusive evidence that people don't intermittently consume therapeutic doses of lithium in those countries. The distribution of lithium concentration in food could be so discontinuous that knowing the maximum value out of the nearly 3,000 samples we have from France, Canada and NZ doesn't give us much information about what obese people in those countries are likely to have encountered in their lifetime. But the hypothesis that people are intermittently consuming doses that high (at least in those countries) would have to do a *lot* of work in order to be consistent with our observations.

## 30 mg/day is not a relevant cutoff

In addition to exclusively mentioning studies with unusual findings in their literature review, the SMTM authors use a subtle sleight of hand to argue that intermittent exposure to therapeutic doses of lithium through food could be the cause of the obesity epidemic. They first say that lithium therapy causes weight gain, citing [Vendsborg et al. \(1976\)](#), in which the average daily elemental lithium dose was 200 mg (SD 8 mg), and also citing [this review paper](#), in which the lowest serum lithium concentration of any patient seems to have been 0.45 mEq/L. Then, later, they claim that the lower end of the therapeutic range of lithium is 30 mg/day, citing a guy on Reddit saying that he takes that much and also has bipolar II, and then they conclude their argument by saying that food could plausibly sometimes contain 30 mg doses of lithium per serving (again, using studies with unusual findings as evidence for this last part).

I say this is a "subtle sleight of hand" because 30 mg/day has doubtful usefulness as a cutoff for a therapeutic dose. As far as I can tell, *no* patients in the weight gain studies cited by the SMTM authors seem to have been taking a dose as low as 30 mg/day.

Moreover, [every \(a\) source \(a\) tells \(a\) you \(a\) that \(a\) the \(a\) therapeutic \(a\) range \(a\) of \(a\) serum \(a\) lithium \(a\)](#) concentration starts at  $\geq 0.4$  mEq/L (most often  $\geq 0.6$  mEq/L) (seriously, just [Google "lithium therapeutic range"](#), the same numbers are everywhere). And unless you're  $>65$  or have really bad kidneys, 30 mg/day [is not enough](#) to get you there - the average adult between the ages of 20 and 65 needs more than three times as much (about 100 mg/day of elemental lithium, or 535 mg of lithium carbonate) to reach 0.4 mEq/L. People over 65 and those with unusually bad kidneys need lower doses, but that hardly helps to explain the obesity epidemic, since [people are much more likely to gain weight when they're young](#). (Yes, older people do tend to be fatter (up to a certain age) but the *rate of change* of excess weight is highest in young adulthood.)

Similarly, Googling "lithium dose range" reveals that [every single website](#) says that the lowest daily dose is at least 600 mg of lithium carbonate (112.2 mg of elemental

lithium), with [some](#) ([a](#)) [sources](#) ([a](#)) saying it is as high as 900 mg (168 mg of elemental lithium).

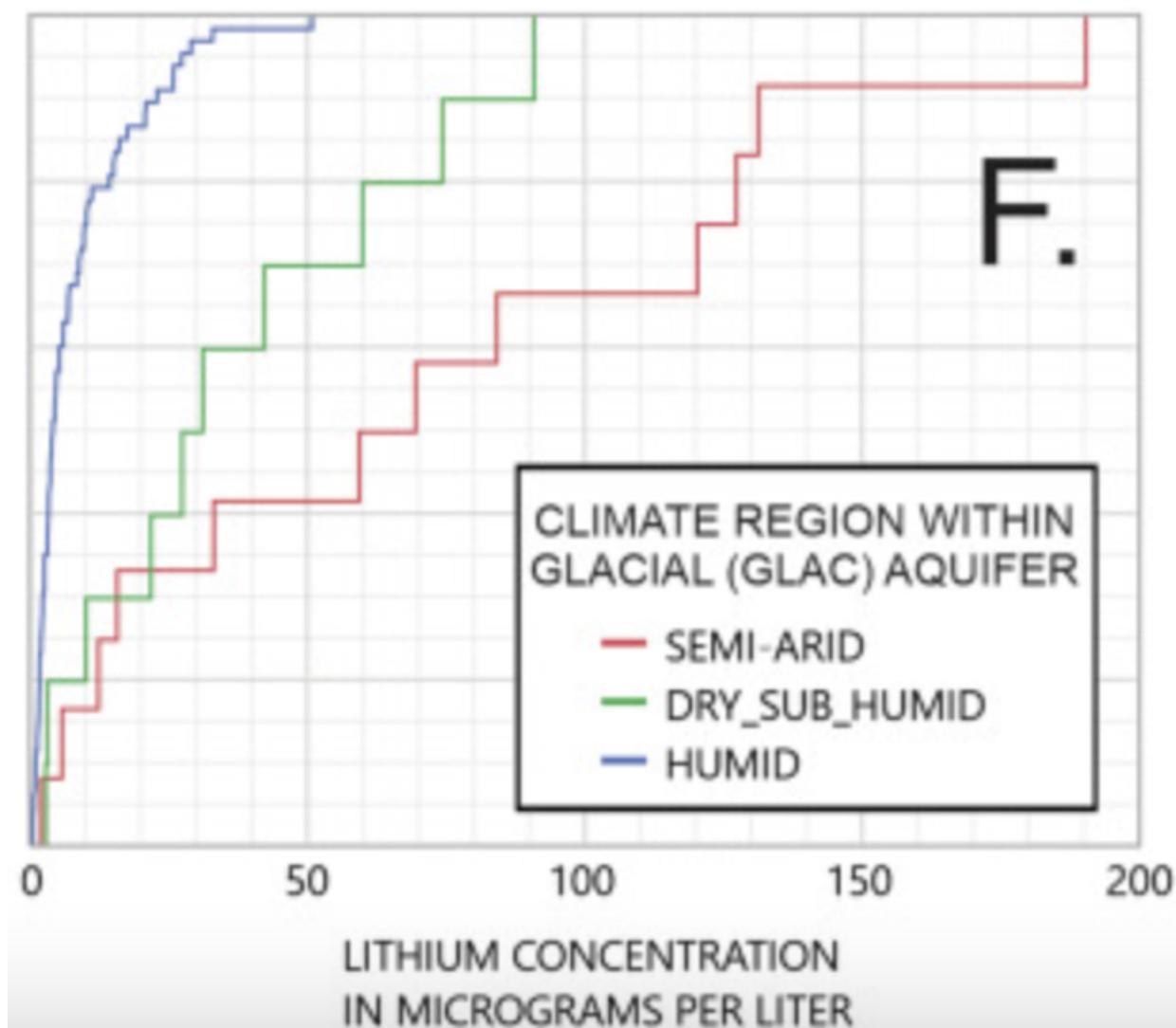
In practice, patient compliance with psychiatric treatment is often not perfect, so there *are* some people with lower lithium serum concentrations [in some studies](#). However, studies with patients who have low average serum lithium concentrations – under 0.6 mEq/L – [seem to find only very modest increases in body weight](#) (or even body weight decreases for patients switching from higher doses)<sup>[5]</sup>. Moreover, patients with serum lithium concentrations that low [experience higher relapse rates](#), so doctors tend to aim for higher serum lithium concentrations when toxicity is not an issue. The SMTM authors do not address any of that, but rather consistently guide the reader into thinking that a 30 mg/day dose can be expected to lead to therapeutic benefit and weight gain.

## (Briefly) Lithium in air and water

The Canadian government has a [page](#) ([a](#)) with a lot of information about lithium, including concentrations found in the air. From the looks of it, if you breathe outdoor air at the 95th percentile of lithium concentration in Canada, by doing so you will consume 2.86 nanograms of lithium per day. This number is really low, and the numbers for indoor air are even lower, so we probably shouldn't worry about this. [The data we have for the US](#) is similar (though unfortunately, it's a lot older).

As the SMTM authors have [gone into](#) ([a](#)), the USGS [has measured](#) ([a](#)) the lithium concentration of thousands of samples of groundwater in the United States, finding an average of 19.7 µg/L, a median of 6.9 µg/L and a maximum of 1.7 mg/L across 3,140 samples of groundwater from used wells. Values tend to be similar or lower in other countries, except for Chile and Argentina, where they're much higher.

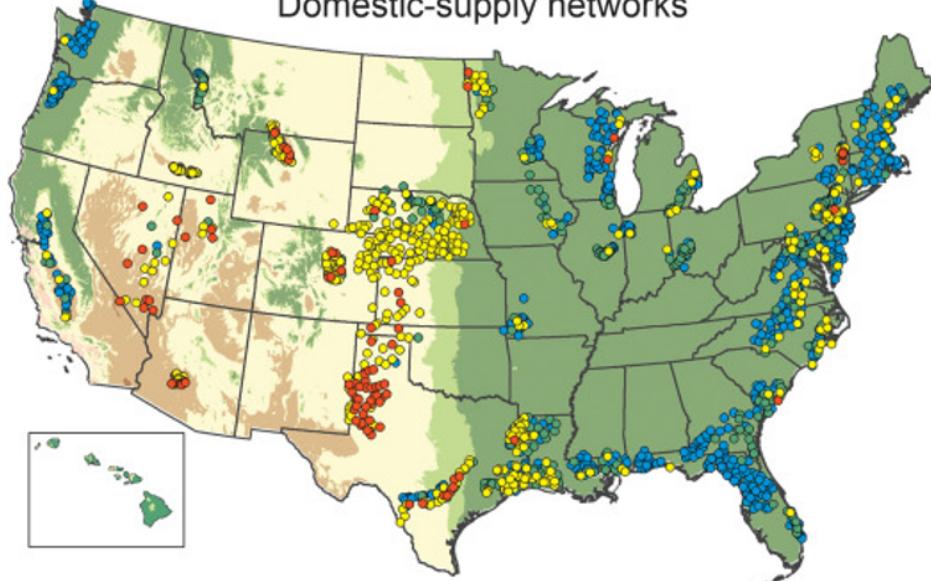
Importantly, lithium concentrations vary a lot according to lithology and climate, so it seems implausible that people in e.g. humid regions are consuming crazy doses of lithium from their water, as this figure displaying the cumulative distribution of lithium concentrations by climate demonstrates:



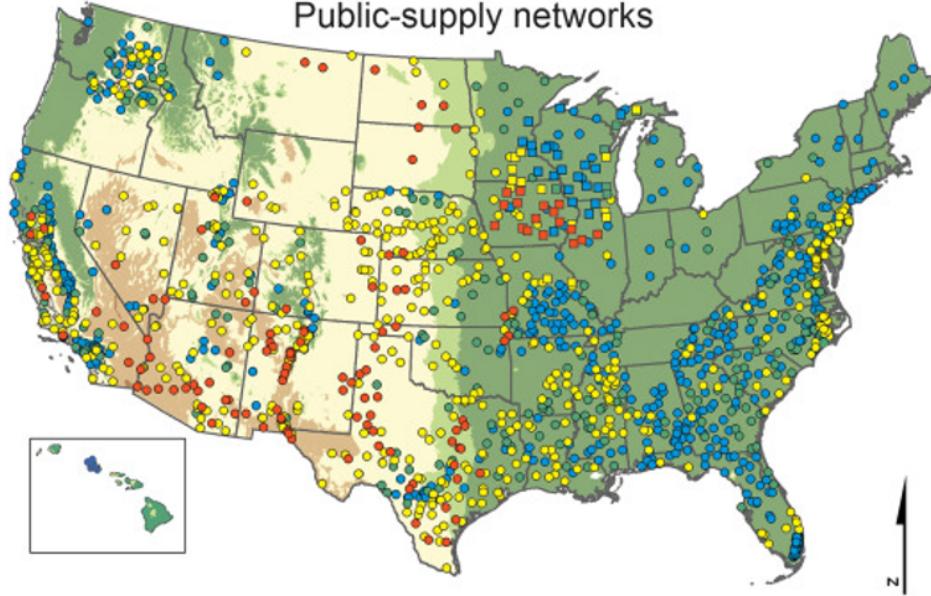
(Source: Figure 2 from [this paper](#).)

And this is important, because a lot of the highest-obesity states, such as those in the South, are in humid regions:

### Domestic-supply networks



### Public-supply networks



### EXPLANATION

0 250 500 1,000 Kilometers

- | Well and Lithium Concentration in $\mu\text{g/L}$   | Climate Region      |
|---|---------------------|
| ● < 5   | Humid               |
| ● > 5 and < 10                                      | Dry Sub-humid       |
| ● > 10 and < 60                                     | Semi-arid           |
| ● > 60  | Arid and Hyper Arid |
| □ Public-supply well in Cambrian-Ordovician aquifer |                     |

Even not taking that into account, however, it seems implausible that a large fraction of the US population is getting therapeutic doses of lithium from their groundwater — even intermittently — given that the maximum concentration across 3,140 samples was found to be only 1.7 mg/L, and given that the average person has only lived for about 15,000 days.

## **Serum lithium concentration data, just like food data, is strong evidence against the hypothesis that people are exposed to high doses of lithium**

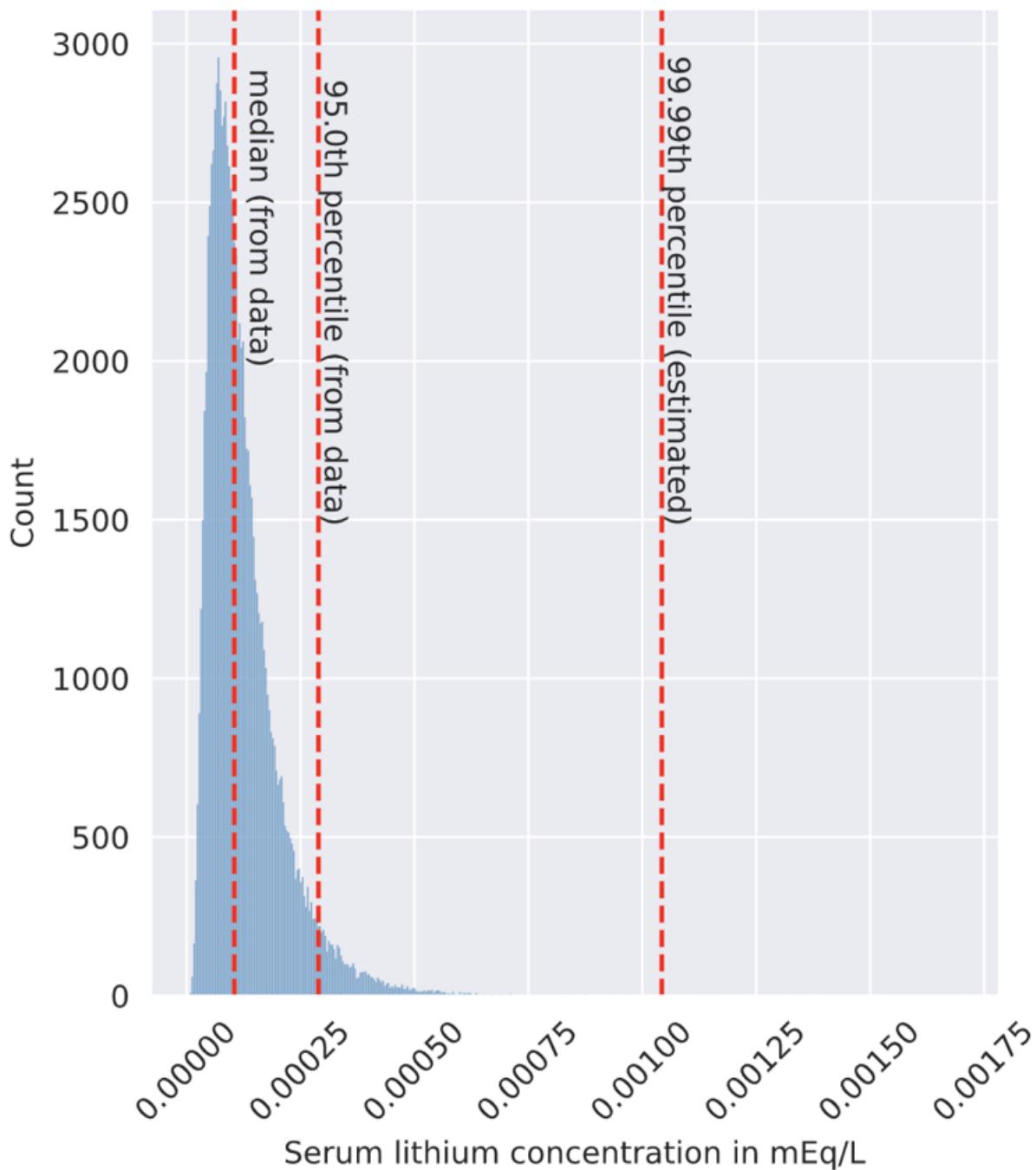
First, I'll give some context for this section. [The therapeutic range of serum lithium concentration in adults is 0.4 to 1.2 mEq/L, when measured at the trough, and a higher rate of relapse is described in subjects maintained at <0.6 mEq/L](#).

The data we have for serum lithium concentration in the general population is probably measured at the trough too (i.e. early in the morning, before people have had any water or food). So with some caveats, [\[6\]](#) it probably makes sense to compare those values directly.

---

The Canadian Health Measures Survey [found](#) that the median whole blood lithium concentration was 0.000068 mEq/L in a nationally representative sample of 5,752 subjects, with data collected from 2009 to 2011. Whole blood lithium concentrations are [65% the value of plasma concentrations \(a\)](#), which are in turn [similar to those in serum](#), so that corresponds to a serum concentration of 0.0001 mEq/L. This value is 4,000 times lower than the lower end of the therapeutic range, and the 95th percentile (0.00029 mEq/L) is 1,383 times lower.

Given that we have two percentiles, we can estimate the parameters of the underlying distribution, and thus its mean and (e.g.) 99.99th percentile, if we assume that it has a certain shape. Under the assumption that the distribution is lognormal, the average and 99.99th percentile of serum lithium concentrations in that Canadian study are 0.00013 mEq/L and 0.001 mEq/L – respectively, 3,000 times and 400 times lower than the low end of the therapeutic range. This is what the distribution looks like, if we sample from it 100,000 times:



(Feel free to look at my code in [this Google Colab notebook](#).)

If we assume that serum lithium concentration is *Pareto-distributed* instead, then the 99.99th percentile is 0.0045 mEq/L, still 89 times lower than the low end of the therapeutic range.

The second-largest study of serum lithium concentrations in the general population that I found was [this one](#) ( $N = 928$ , data collected around 2010 in Germany), in which the median is 0.000138 mEq/L and the 91.7th percentile is 0.0003 mEq/L. Plugging those numbers in the same code, we find a 99.99th percentile of 0.0011 mEq/L in a lognormal distribution and 0.0055 mEq/L in a Pareto distribution, the latter of which is still 73 times lower than a therapeutic dose.

---

Some places have much higher lithium exposure. In northern Chile, where the highest drinking water concentrations of lithium in the world were found in the 1970s, serum concentrations [were](#) only about two orders of magnitude lower than the therapeutic range; [a more recent study](#) in nearby northern Argentina, which likewise has extremely high levels of lithium in its rivers, found similar (but a bit lower) levels. And, as we've gone into before, some older studies from Germany suggest that lithium levels were quite high in food at the time the measurements were made. Moreover, in the Canary Islands, average dietary lithium consumption [has been estimated](#) at 3.7 mg/day.

But I think it's important to point out this data from Canada, France, New Zealand, etc., because the obesity epidemic is a pretty global problem that has definitely reached those countries, and that is [worse](#) in Canada (31% obese in 2016) and NZ (32%) than in Chile (29%), Argentina (29%) [or the Canary Islands \(20.1%\)](#), despite the first two countries having much lower average lithium exposure.

## Clinical doses of lithium cause a lot more side effects than just weight gain

But let us suppose for a moment that people *are* inadvertently taking therapeutic doses of lithium from their food/air/water/whatever every few months or years. There's still another problem with the lithium hypothesis: why aren't people getting the *other* lithium side effects?

### Hand tremors

[This Cochrane review](#) found that lithium greatly increased the incidence of hand tremors (OR 3.25, 95% CI 2.10 to 5.04;  $N = 1241$ ;  $k = 6$ ) among patients taking the drug for a few weeks to control acute mania.

### Hypothyroidism

Hypothyroidism is [about six times more common in patients on lithium](#). And notably, its prevalence through time and space doesn't seem to follow the pattern of obesity. It [hasn't](#) been consistently becoming more common over time, and global rates of the disease [don't seem correlated with obesity rates](#), with thin countries like China and Brazil having higher prevalences. Moreover, it is more common in old age, whereas weight gain is more common in youth.

### Diabetes insipidus

But perhaps the most specific side-effect of clinical doses of lithium is acquired nephrogenic [diabetes insipidus](#).

This disease is not well-known, so I'll explain what it is. Diabetes insipidus (henceforth DI) is a disease characterized by polyuria (peeing a lot), polydipsia (increased thirst and fluid intake), and abnormally low urine concentration. Polyuria and polydipsia are also found in untreated diabetes mellitus, hence the similar name, but DI has nothing to do with blood sugar.

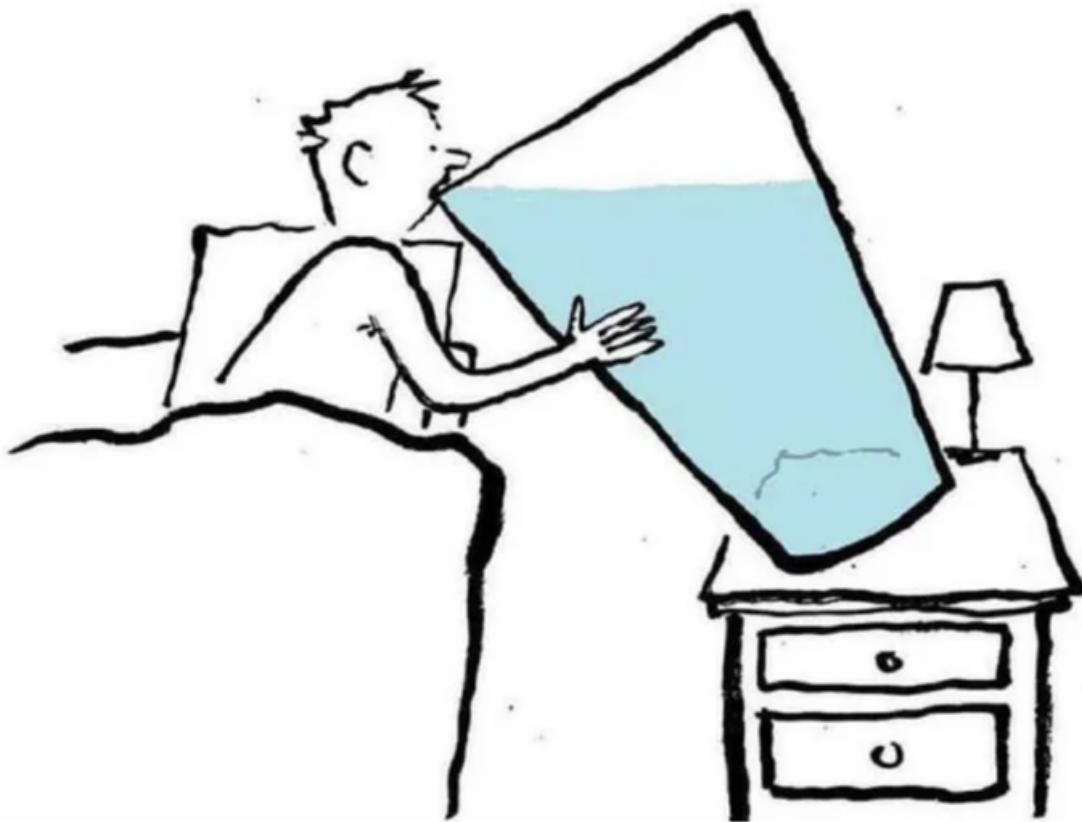
Most cases of DI in the general population are caused by abnormally low secretion of [vasopressin](#) (aka antidiuretic hormone, which gives the signal to concentrate urine); this type is called "neurogenic" or "central." The type of DI that lithium causes is called "nephrogenic," and is instead caused by the kidneys becoming unable to respond to vasopressin. Someone with either of those types will pee out a lot of clear urine even after a long period of fluid deprivation — but those with neurogenic DI function normally if they are given desmopressin, so it's easy to tell those types apart.

Nephrogenic DI seems to be extremely rare in the general population. The NHS website [says \(a\)](#) that the prevalence of any kind of DI is 1 in 25,000, and that DI is nephrogenic in nature in "rare cases." Acquired (non-hereditary) nephrogenic DI is so rare that lithium therapy is literally [its most frequent cause](#).

How common is diabetes insipidus in patients on lithium therapy? In [Vendsborg et al. \(1976\)](#), 30% of patients had it. Subclinical symptoms of the disease, such as abnormal urine concentration ability, are more common, affecting 54% of 1,105 unselected patients in [this study](#). Notably, the second top post of all time on the r/Lithium subreddit [is this \(a\)](#):

Nobody:

Me at 3 AM:



Those symptoms can set in early on with lithium treatment – in [Forrest et al. \(1974\)](#), maximum urine concentration was much lower after 8-12 weeks on lithium than before (Cohen's  $d = 3.84$ , an extremely huge effect size), and such impairment has been found after four weeks of lithium treatment [in rats](#).

---

It's true that we don't know to what extent those side effects happen if you take high doses of lithium intermittently rather than chronically, but note that *the exact same argument applies to weight gain*.

### **Weight gain is associated with other lithium side-effects**

Moreover, there's evidence that diabetes insipidus and hypothyroidism might play a role in the weight gain caused by lithium, making it unlikely that lithium exposure in the general population would cause a lot of weight gain *in the absence of those other*

side effects. In [Vendsborg et al. \(1976\)](#), weight gain was much greater among those with greatly increased thirst than among those without:

**Table 4. Weight gain after start of lithium treatment in 62 patients with increased thirst**

Thirst	Patients		Recorded weight gain (kg)
	Number	Percentage	
Little increase	8	13 %	1.9 ± 0.67*
Moderate increase	15	24 %	5.3 ± 1.6
Great increase	39	63 %	7.0 ± 0.98

\* =  $P < 0.01$  (versus moderate and great).

and much greater among those with clinical diabetes insipidus than among those without:

**Table 5. Weight gain and lithium metabolism in patients without and with clinical diabetes insipidus**

	Weight gain kg		Lithium dose (mmol/24 h)	“Lithium clearance” (ml/min)
	Case record (n = 70)	Questionnaire (n = 45)		
Without clinical diabetes insipidus	4.7 ± 0.7	4.2 ± 0.8	29.2 ± 1.2	36.0 ± 1.9
With clinical diabetes insipidus	8.3 ± 1.4	11.8 ± 1.5	34.6 ± 2.1	43.8 ± 3.3
<i>P</i> value	<i>P</i> < 0.01	<i>P</i> < 0.001	<i>P</i> < 0.02	<i>P</i> < 0.05

[Vestergaard et al. \(1980\)](#) report a similar finding. Those studies note that increased thirst can cause weight gain by increasing the consumption of caloric drinks.

On top of that, weight gain is [a well-known symptom of hypothyroidism](#).

(Note: I have attempted to make some of those points in the comment section of [SMTM's last post about lithium](#), but they never approved my comment. (They did approve some comments made after mine.) [There is no mention of diabetes insipidus anywhere on SMTM's website](#).)

## Even therapeutic doses of lithium don't cause enough weight gain to

# explain the obesity epidemic

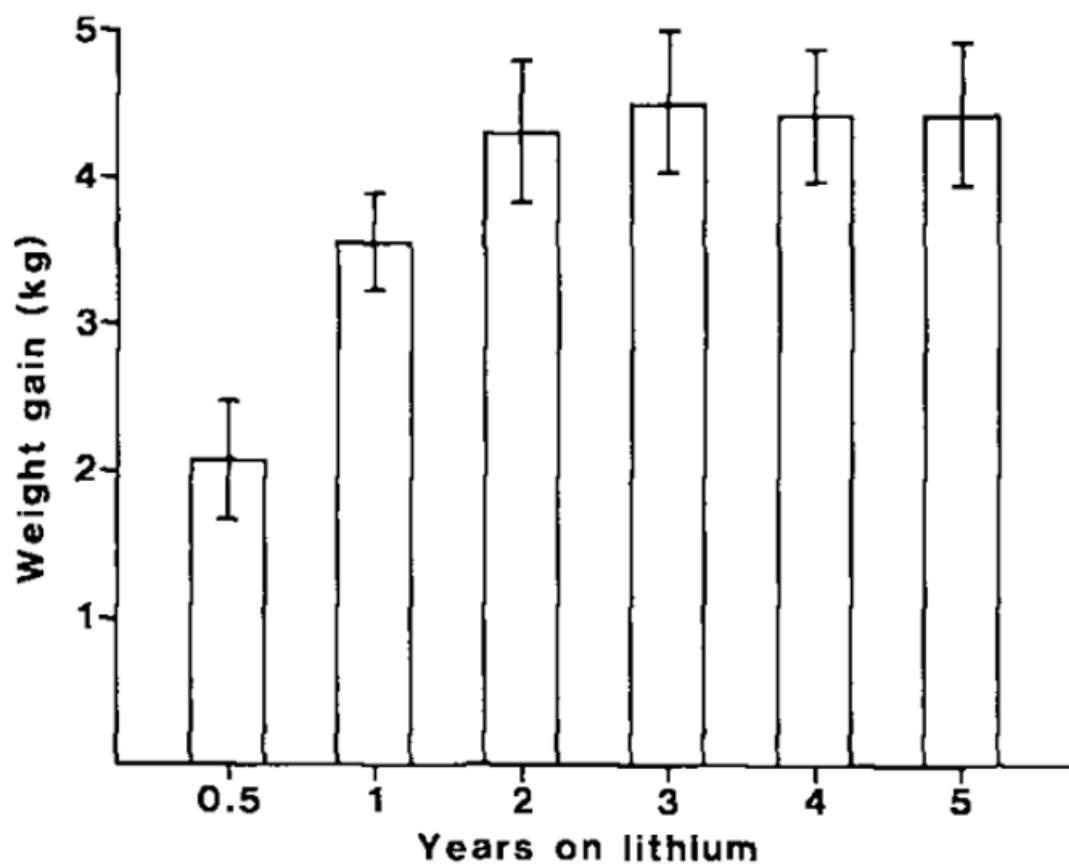
How much weight gain does lithium even cause, on average? The SMTM authors have cited figures from [Vendsborg et al. \(1976\)](#) and [Vestergaard et al. \(1980\)](#), and I found a number of others ([Versteegard et al. \(1988\)](#), [Mathew et al. \(1989\)](#) and [Armond \(1996\)](#)) in my own research:

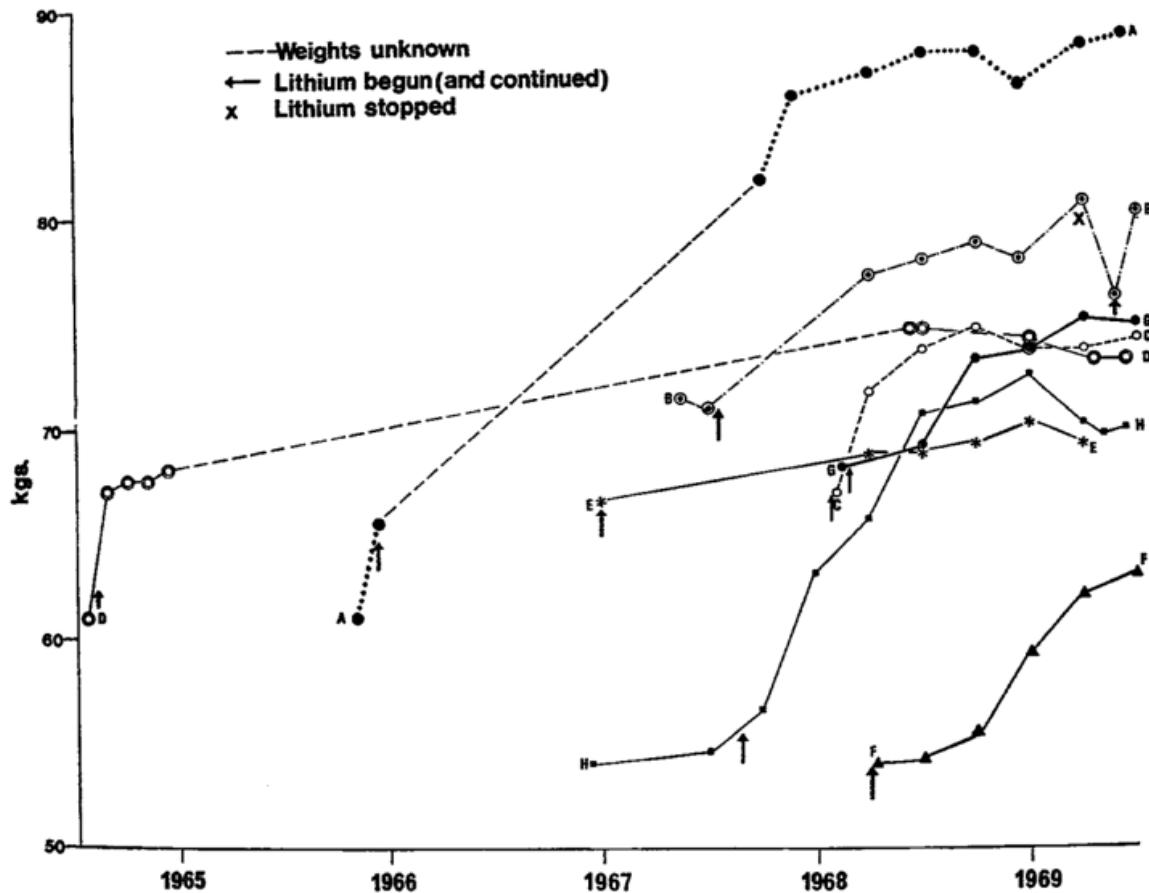
Source	Duration of treatment	Weight gain	Average weight/BMI	Daily dose	Serum concentration
Vendsborg et al. (1976)	7 years	5.9 kg		200 ± 8 mg	
Vestergaard et al. (1980)	5.2 years	20% >10 kg	76 kg	226 (83 - 395) mg	0.85 (0.2 - 2.0) mEq/L
Versteegard et al. (1988)	1.83 years	4 kg	70.4 kg	161 mg	0.68 mEq/L
Mathew et al. (1989)	4.7 years	1 kg/m <sup>2</sup> <sup>[1]</sup>	28 kg/m <sup>2</sup>	129 mg	0.54 (0.2 - 1.6) mEq/L
Armond (1996)	15.5 years	0.35 kg <sup>[2]</sup>	79 kg		

[1] Change in average BMI. [2] versus controls.

All of those studies are observational and none other than the last one has a control group. Patients with bipolar disorder often take antipsychotics and antidepressants in addition to mood stabilizers such as lithium, and the patients in these studies are no exception. In the first two studies, for instance, only a minority of patients were taking lithium alone. Antipsychotics cause weight gain, and the same is true of several antidepressants that were available when those studies were conducted (e.g. TCAs and the MAOI Nardil), so lithium alone probably causes less weight gain than those numbers suggest.

Weight gain [does not seem](#) to be constant throughout the duration of lithium treatment; it instead slows down or even stops at some point, as the figures below, from [Versteegard et al. \(1988\)](#) and [Kerry et al. \(1970\)](#), respectively, illustrate:





#### WEIGHT CHANGES IN LITHIUM RESPONDERS

Cases A, D, G, H, & F are females

Cases B, C, & E are males

I mentioned these studies despite their methodological flaws because they have the longest follow-up times I've ever found in the literature (and because the SMTM authors cited some of them). What do meta-analyses of RCTs have to say?

I searched Embase for systematic reviews and meta-analyses on the effects of lithium on weight gain, [7] and found 3. [This one, from 2012](#), finds that patients on lithium are almost twice as likely to experience clinically significant weight gain (defined as gaining  $\geq 7\%$  of your body weight) as patients on placebo ( $k$  (number of studies) = 5). The two studies that reported patients' serum lithium concentrations found values of [0.8  \$\pm\$  0.3 mEq/L](#) and [0.66  \$\pm\$  0.27 mEq/L](#); the other studies said that patients were expected to have serum concentrations of at least 0.8 mEq/L.

The second result was a [Cochrane Review](#), which found that people on lithium for acute mania were a bit more likely to gain weight (OR 1.48, 95% CI 0.56 to 3.92;  $n = 735$ ,  $k = 3$ ); though the reviewers say that there is "insufficient evidence" that lithium had any effect. This review only included patients in the mental hospital for mania, however, so the average duration of treatment was probably really short.

Oddly, the third result I found was a [2022 systematic review and meta-analysis](#) that says that weight gain during lithium treatment is not statistically significant from zero, and is significantly greater in *shorter* studies than in longer ones ( $k = 9$ ,  $n = 991$ ). Compared to placebo, the meta-analysis finds that lithium causes less weight gain ( $k = 3$ ,  $n = 437$ ). I don't buy this paper's conclusion, but I think this is nonzero evidence that lithium causes somewhat less weight gain than some other studies suggest.

So lithium seems to cause an average of zero to 6 kg of weight gain in the long term. And strikingly, the upper end of that range, although large, is only half the amount of weight the average American adult has gained since the early 70s, which, according to my analysis of NHANES data (which is all in [a public Google Colab notebook](#)) is about 12 kg (~26.4 lb), and as high as 15.7 kg (~34.5 lb) for people in their 30s. (The *median* American adult gained about the same amount of weight.)

And it's not as if Americans were that thin in the early 1970s! [47% of adults were overweight and 14.5% obese](#) ([a](#)). In contrast, obesity rates [are under 3%](#) in traditional societies that engage in foraging or subsistence farming. Moreover, there is [substantial](#) ([a](#)) [evidence](#) ([a](#)) that Americans gained a lot of weight before 1970. It's hard to know the overweight and obesity rates of the general population back in the 19th century, because there was no NHANES back then, but we do know that [men at elite colleges](#) ([a](#)) ([source](#)), [Citadel cadets](#) ([a](#)) ([source](#)) and [veterans](#) all started getting substantially fatter in the early 20th century.

I feel the need to stress this, because the SMTM authors [claim](#) ([a](#)) that there was an abrupt shift in obesity rates in the late 20th century, a claim that is [probably based to some extent on an artifact of the definition of BMI](#) ([a](#)), and so some people reading this might have the impression that 1970s Americans were really thin or something, when they *really* weren't.

Anyway, so the 12 kg average weight gain since the early 70s is not the whole story. A 5'9" man with a BMI of 21, which is higher than the average [for the Hadza](#) (independently of gender or age), for Citadel cadets born in the 1870s, and for men entering Amherst, Yale, or Harvard in the 19th century, is 19.5 kg (43 lb) lighter than the average college-aged American man today. Moreover, a 5'9" man with the average BMI of 19th-century veterans in their 40s is a whopping 22.5 kg (49.5 lb) lighter than the average American man of the same age and height today.

So even if everyone in the US were consuming therapeutic doses of lithium without knowing it, that would leave most of the secular increase in body weight unexplained.

---

This can also be seen in studies that report the obesity rate or average BMI of patients taking lithium back when obesity rates were low. [Chen & Silverstone \(1990\)](#) (a review article that has been cited by SMTM) reviewed some studies reporting either of those figures, and in *none* of them was the obesity rate greater than 25% - even though most of those patients were probably on antipsychotics and/or TCAs as well. So it's difficult to imagine how lithium exposure could explain why the obesity rate is greater than 30% in several countries.

## Lithium weight gain seems to (perhaps) be dose-dependent even at

# therapeutic doses

Furthermore, there's some evidence that weight gain on lithium is dose-dependent even at therapeutic doses.

[Gelenberg et al. \(1989\)](#) randomized 94 patients on lithium therapy to either a normal or a low dose in a double-blind trial. They found that patients on the low-dose group were about half as likely to report worsening weight gain.

[Abou-Saleh and Coppen \(1989\)](#) likewise randomized 91 patients on lithium therapy to either maintain their dosage or decrease it by up to 50%. The group with the lowest dosage lost 0.9 kg, whereas the highest-dosage group gained the same amount of weight.

[Keller et al. \(1992\)](#) performed the same kind of study, and found that patients in the normal-dose group were more likely to report weight gain than those in the low-dose group.

Moreover, in [Vendsborg et al. \(1976\)](#), one of the observational studies we've talked about, there was a 0.44 correlation between lithium dosage and weight gain.

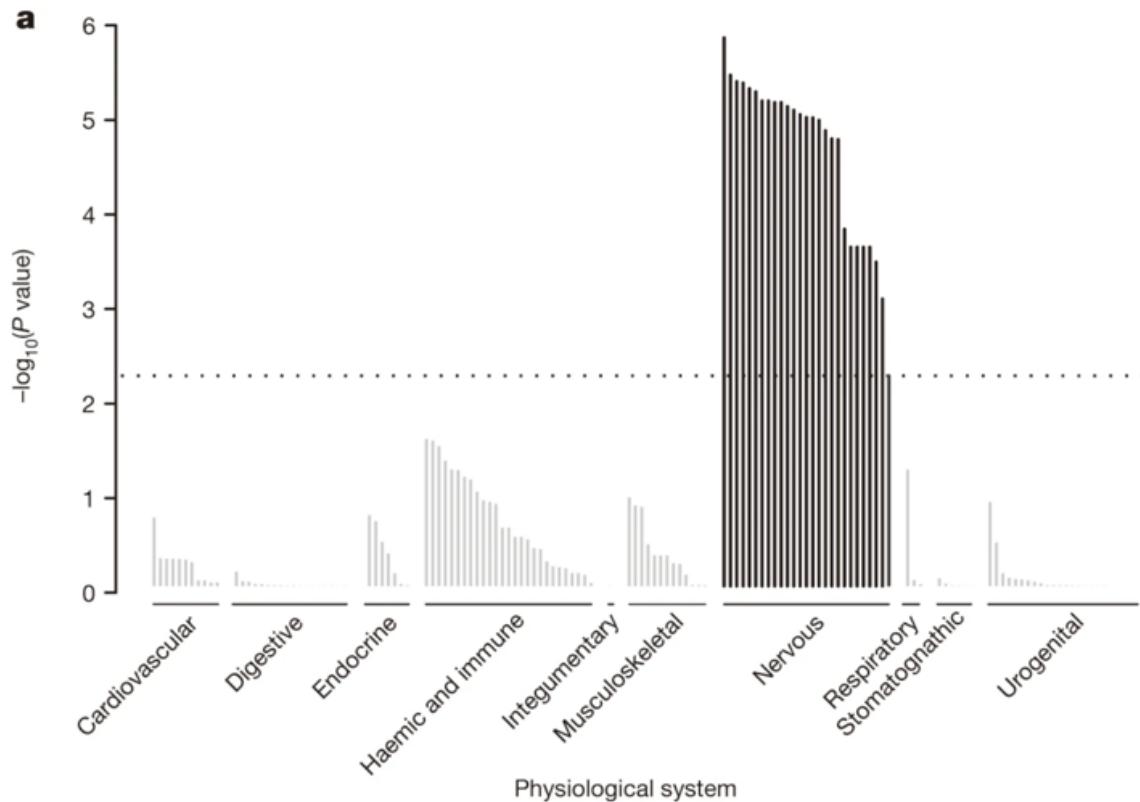
Not all studies find such a dose-dependence - [Mathew et al. \(1989\)](#) did not find one. But the patients in this study had a low serum concentration (0.54 mEq/L on average) and they barely gained any weight (as I've mentioned, their average BMI increased by 1 kg/m<sup>2</sup> over the course of 4.7 years). [Vestergaard et al. \(1980\)](#) and [\(1988\)](#) do not find it either. So this seems to be something that might exist, but we're not sure.

It seems noteworthy, however, that the studies with the best methodology (the first three studies I mentioned, which were randomized double-blind trials) all found dose-dependence. Moreover, when scientific papers say that some association has not been found, often what they mean is that it just hasn't reached an arbitrary significance threshold. Since [Mathew et al. \(1989\)](#), [Vestergaard et al. \(1980\)](#) and [\(1988\)](#) do not provide any actual data on the amount of weight gained by patients as a function of dosage, for all we know that could be what is going on.

Somewhat relatedly, [Rinker et al. \(2020\)](#) randomized 23 patients with multiple sclerosis to take either placebo or a low dose of lithium (150-300 mg/day of lithium carbonate, ~30-60 mg/day of elemental lithium) for a year in a crossover trial. 13 patients complained of weight gain, but the other 10 complained of weight loss.

## Genes that influence BMI do not tend to be expressed in the kidneys (which govern lithium secretion)

If the obesity epidemic were caused by lithium, we should probably expect poor kidney function to predict weight gain, since it [strongly predicts](#) serum lithium concentration. But the genes that affect obesity are [primarily expressed](#) in the nervous system, and the urogenital system doesn't stand out at all:



## The evidence that trace doses of lithium exert significant effects is actually pretty weak

The SMTM authors [say \(a\)](#):

There's lots of evidence (or at least, lots of papers) showing psychiatric effects [of lithium] at exposures of less than 1 mg [...]. If psychiatric effects kick in at less than 1 mg per day, then it seems possible that the weight gain effect would also kick in at less than 1 mg.

I'd like to point out that Gwern has [looked into this \(a\)](#) and concluded that the evidence that such low doses of lithium cause psychiatric effects is actually fairly weak.

## Mysteries that lithium cannot explain

A *Chemical Hunger* [opens up](#) with a list of mysteries related to the obesity epidemic. I don't endorse the list – for one, “wild animals are becoming obese” seems to have been pretty much made up (see the fourth point in [this comment](#)), and all evidence we have that “lab animals are becoming obese” is exactly one (1) unreplicated paper co-authored by a guy that has been involved in numerous controversies regarding his conflict of interest with the processed food and restaurant industries.<sup>[8]</sup>

However, some of the mysteries *are* genuinely true and interesting – for instance, that obesity rates are much lower in high-altitude areas, and that human food is unusually palatable. And lithium does not seem to explain either of those mysteries.

(Note that a theory of the obesity epidemic *does not need* to explain these mysteries. They could very well be a result of factors completely unrelated to the secular increase in BMI since the early 20th century. But I feel like some people might believe that the lithium hypothesis neatly explains all of SMTM's mysteries, when it doesn't.)

## Altitude

Using [publicly-available data from the USGS](#) and the [Open Elevation API](#), I found that across 1,027 domestic-supply wells (all wells whose coordinates were available), the correlation between altitude and  $\log(\text{lithium concentration})$  is 0.46. I also checked the correlation between altitude and *topsoil*  $\log(\text{lithium concentration})$  in the United States, with data I found [here](#), and, again, it was positive (0.3). So lithium exposure is probably *higher*, rather than lower, in high-altitude areas in the United States (which, as a reminder, have *lower* obesity rates).

## Palatable human food

To the extent that the mystery is that human food is unusually *palatable*, and not *just* that it's unusually fattening, it's hard to see how that could be explained by lithium — or any other contaminant, for that matter — being in the food, unless the contaminant happens to be tasty.

## Youth (Not mentioned in their blog post series)

As I mentioned before, people gain excess weight much more rapidly when they're young, [at all BMI levels](#). I think this deserves to be deemed a mystery. And, importantly, it is the *opposite* of what you would expect if lithium were the cause of the obesity epidemic, because among patients on lithium therapy, controlling for their dosage, [young people have lower serum lithium concentrations](#). In the general population, too, it has been found both in [Canada](#) and [Germany](#) that serum lithium concentration increases with age.

This doesn't make the lithium hypothesis *impossible* – weight gain might be primarily determined by lithium consumption or throughput rather than by serum concentration at any given time. But I think it's still clearly the case that a positive association between advanced age and weight gain is more likely in worlds in which the lithium hypothesis is true than in worlds in which it is false, and given that the association *goes the other way* in our world, we should update against the lithium hypothesis accordingly.

## Conclusion, bets and bounties

When I imagine a world in which the obesity epidemic is caused by environmental lithium exposure, this is what I expect it to look like:

- some places in the Andes and the Canary Islands have way higher obesity rates than everywhere else,
- Vietnam has an obesity rate closer to that of New Zealand,
- old people are more likely to gain excess weight than young adults,
- there is a nephrogenic diabetes insipidus epidemic,
- unexplained hand tremors are common,
- genes that affect obesity are disproportionately expressed in the kidneys (in addition to the central nervous system),
- hypothyroidism rates are going up,
- countries and age groups with higher hypothyroidism rates also have higher obesity rates,
- therapeutic doses of lithium cause at least ~20 kg of weight gain *on average*, and
- people with chronic kidney disease sometimes mysteriously die of lithium toxicity.

As far as I can tell, none of those things are true. So my credence that lithium exposure plays a major role in explaining the obesity epidemic is very low, something like 0.1%.

## Bets

Several months ago, my husband [publicly challenged](#) the SMTM authors to a bet on their contamination theory of obesity. They have declined to bet. I'd like to remind them that the bet offer is still active.

## Bounties

I am offering a \$40 bounty for each Metaculus or Manifold Markets question about the contamination theory of obesity that both I and the SMTM authors agree to be a good test of some aspect of the theory. I'll pay for up to 5 questions. This bounty expires 90 days after the publication of this post.

I am also offering a \$300 bounty for anyone who writes a comment convincingly arguing that the lithium concentration data from the large and recent studies I found from France, Canada, Italy and New Zealand isn't a good indication of how much lithium people in those countries get from their food, and that it's actually quite likely that the average dietary intake in those countries is closer to 1 mg/day for a normal-sized adult. Pointing out minor caveats to the interpretation of those studies would not count (though it's definitely welcome), you have to argue that the data I found is very weak evidence that dietary lithium consumption is on the order of 10-50 µg/day instead of 1 mg/day in those countries. This bounty expires 90 days after the publication of this post.

Right now, whether and to whom this last bounty is paid out is fully up to my judgment. But I am also offering a \$50 meta-bounty for whoever comes up with better, objective criteria for that bounty. This meta-bounty expires in 90 days, and if it is fulfilled, then the \$300 bounty will expire 90 days after the fulfillment of the meta-bounty.

Update: Austin Chen [has offered to match these bounties](#) (in [Manifold Markets](#) currency).

## Acknowledgments

Thanks to Holly Elmore, Katherine Worden, Philipp Risius, [@Willyintheworld](#) and my husband Matthew Barnett for helpful comments and suggestions in earlier drafts of this post.

I do not speak for anyone other than myself, and all errors are my own.

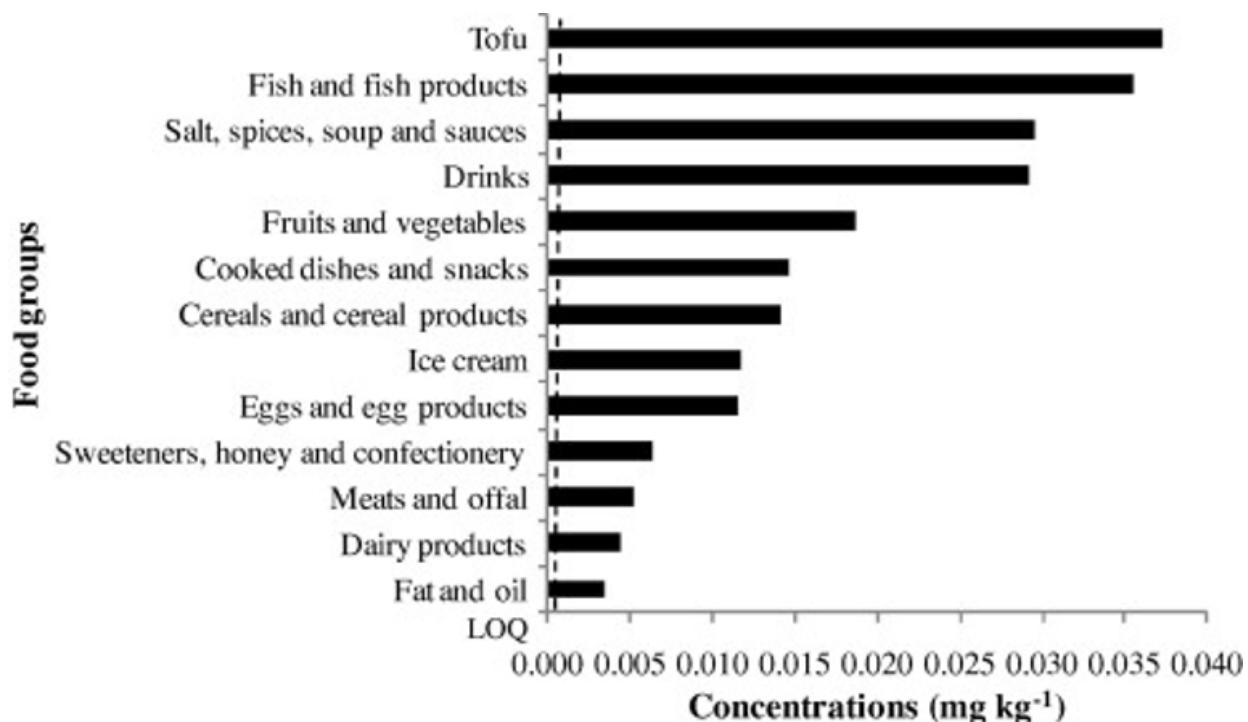
## Addenda

### Animal products don't have that much lithium, according to data from France, Canada and Italy

SMTM found, in their food literature review, that animal foods have quite a bit of lithium. They think that that might explain why vegetarian and vegan diets lead to some weight loss. Quoting from their post:

Pretty much everything we see suggests that animal products contain more lithium on average than plant-based foods. [...] It's interesting, though not surprising, to see such a clear divide between plant and animal foods. In fact, we wonder if this can explain why [vegetarian diets seem to lead to a little weight loss and vegan diets seem to lead to a little more](#), and also why neither of them work great.

However, this horribly fails to replicate in the Total Diet Studies that I've found in my research. For instance, the food item with the highest lithium concentration in [the second French Total Diet Study](#) is... *definitely* not something that vegans are known for not eating:



The [first French Total Diet Study](#) likewise found that meat (2 µg/kg), milk (6 µg/kg) and ultra-fresh dairy products (4 µg/kg), among other animal foods, had a lot less lithium than fruits (7 µg/kg), vegetables (14 µg/kg), miscellaneous cereals (20 µg/kg), nuts and oilseeds (22 µg/kg), and Viennese bread (37 µg/kg), among other plant-based foods.

Moreover, in the [Canadian TDS \(a\)](#), all of the top 10 foods in average lithium concentration are vegan, as are all but 3 of the top 20, as you can see in [this Google Colab notebook](#):

Food Name	count	mean	std	min	25%	50%	75%	max
Salt	3.0	512.576718	535.347138	69.946896	215.066588	360.186279	733.891629	1107.596979
Mineral Water, Carbonated	3.0	280.774141	196.111313	54.777319	218.075647	381.373976	393.772551	406.171127
Herbs & Spices Mixed	3.0	264.661012	101.890624	204.908744	205.836874	206.765004	294.537146	382.309289
Spinach	3.0	148.229592	78.220511	76.269498	106.602961	136.936423	184.209639	231.482855
Fruits, Raisins, Semi-Moist	3.0	136.156375	37.749741	110.897164	114.458773	118.020382	148.785981	179.551580
Fruits, Melon, other varieties	3.0	114.312996	41.463939	67.033002	99.225358	131.417715	137.952992	144.488270
Tomatoes, canned	3.0	109.887639	14.907241	92.754682	104.884645	117.014607	118.454117	119.893626
Baking Powder	3.0	101.224604	20.442450	82.515114	90.315056	98.114998	110.579348	123.043699
Pasta, Mixed Dishes	3.0	73.551374	41.052134	27.280432	57.524724	87.769016	96.686844	105.604672
Soya Sauce	3.0	63.929657	50.867815	26.655680	34.954595	43.253510	82.566645	121.879780
Fast Food, Hamburgers	3.0	58.828846	6.239379	51.780735	56.420108	61.059482	62.352902	63.646323
Bread, Whole Wheat	3.0	57.906364	8.211285	48.426645	55.455198	62.483752	62.646224	62.808696
Cucumbers, English	3.0	49.421010	18.352234	29.181716	41.641686	54.101657	59.540657	64.979658
Condiments	3.0	44.703471	10.841303	32.389845	40.648487	48.907128	50.860284	52.813440
Flour, white (wheat)	3.0	42.822042	5.370180	36.643475	41.049469	45.455463	45.911326	46.367188
Fast Food, Pizza	3.0	42.662551	15.907931	31.835120	33.530405	35.225690	48.076267	60.926844
Muffins, Regular	3.0	41.335168	11.051541	30.927507	35.535714	40.143922	46.538998	52.934074
Bread	3.0	39.051750	4.457283	34.378961	36.949322	39.519682	41.388145	43.256608
Beans, Baked, canned	3.0	37.519319	50.104012	8.302024	8.592239	8.882454	52.127966	95.373478
Cereal, Rice, Bran	3.0	37.222346	28.034386	18.606773	21.100866	23.594960	46.530132	69.465304

In the [large Italian study I have mentioned before](#), the highest lithium concentration was found in fish/seafood (median 19.10 µg/kg) (the same finding as the first French TDS) but legumes (15.43 µg/kg) and cereal (14.83 µg/kg) also had a lot of lithium, whereas meat products (3.41 µg/kg), eggs (3.87 µg/kg) and dairy products (4.78 µg/kg) all had low levels. [This other study from Italy](#) found that fruits and vegetables (33 µg/kg) and cereals and tubers (31 µg/kg) were the food groups with the highest lithium concentrations.

So it doesn't look *at all* like there is a "clear divide" between plant and animal foods.

**We \* do\* have data on the lithium content of processed food**

Again from SMTM's literature review of the lithium content of food:

One thing we *didn't* see much of in this literature review was measurements of the lithium in processed food.

We're very interested in seeing if processing increases lithium. But no one seems to have measured the lithium in a hamburger, let alone a twinkie.

Fortunately, this is incorrect — Canada *has* measured the lithium in hamburgers, and the concentration they found was  $58.8 \pm 6.2 \mu\text{g/kg}$ . They also measured the lithium content in pizza ( $42.66 \pm 16 \mu\text{g/kg}$ ), French fries ( $26.77 \pm 20.8 \mu\text{g/kg}$ ), hot dogs ( $25.49 \pm 3.9 \mu\text{g/kg}$ ), chicken nuggets ( $10.78 \pm 3.4 \mu\text{g/kg}$ ), fried rice ( $12.11 \pm 2.8 \mu\text{g/kg}$ ), prepackaged sandwiches ( $27.14 \pm 1.3 \mu\text{g/kg}$ ), and other types of fast food.

France, too, has measured the lithium concentration of a variety of processed food items, in its first and second Total Diet Studies.

## Factual inaccuracies and misrepresentation of sources in SMTM's posts about lithium

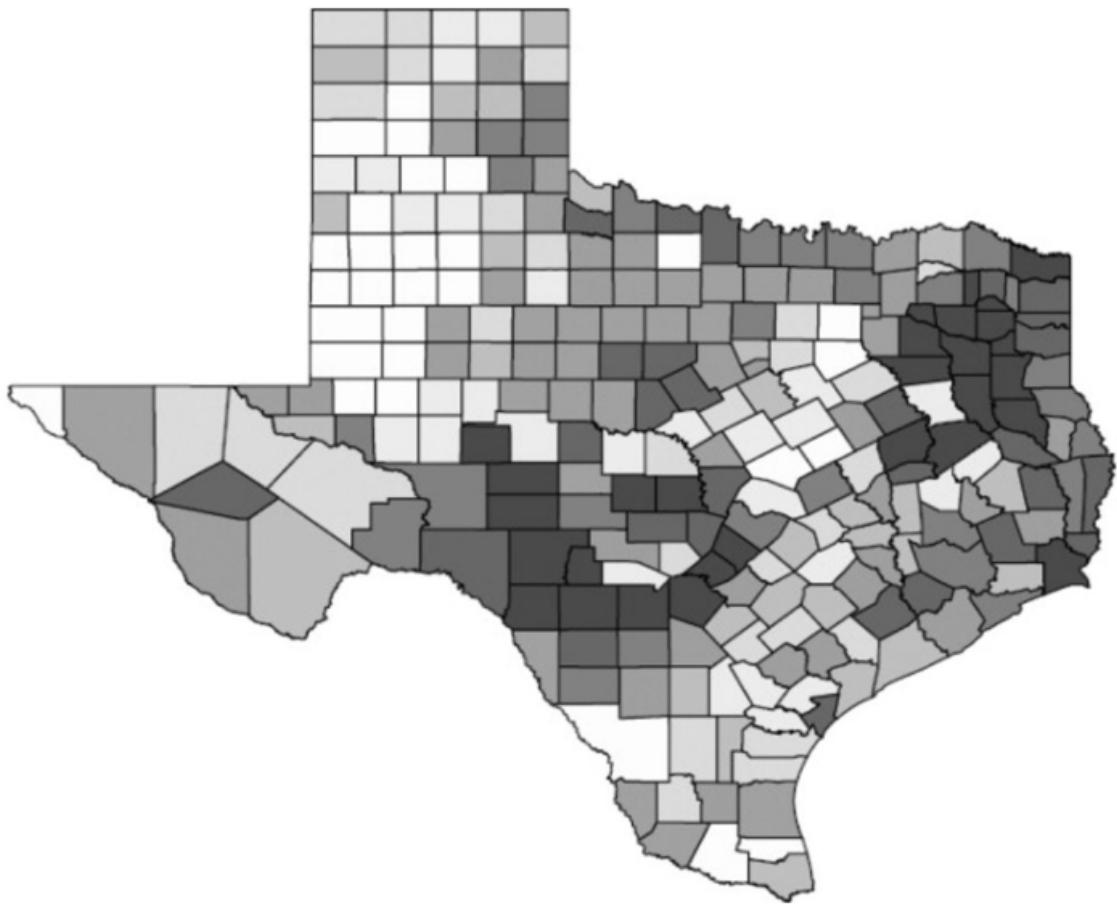
Since we're already here, I decided to point out and correct a few claims that the authors of SMTM have made in posts about lithium that are not supported by their own sources.

### No, Texas counties with higher lithium levels are not more obese

In Part VII: Lithium (a), the SMTM authors say:

In Texas, a survey of mean lithium levels in public wells across 226 counties (Texas has 254 in total) found lithium levels ranging from 2.8 to 219.0 ng/mL. Now Texas is not one of the most obese states — but it tends to be more obese along its border with Louisiana [sic], which is also where the highest levels of lithium were reported.

However, if you look at the source, their map of lithium levels across Texas counties actually says the *opposite* — that counties along the border with Louisiana have *lower* lithium levels than other counties (though it's a bit confusing because on the map darker areas represent lower levels):



Download : [Download full-size image](#)

Fig. 1. Average lithium levels in Texas, 1999–2007 (darker areas represent lower levels).

Someone else has already [pointed this out in the comments of that post](#), several months ago, as have I, [on a 05/08/2022 Twitter thread tagging the SMTM authors](#), but the post has not been fixed, and the authors have not acknowledged or addressed the error in any way.

When you calculate the correlation between log(water lithium levels) and log(obesity %) in Texas, [you find that it is](#) -0.13.

## No, obesity in the West Bank was not 50% in men in 2003

Also from [Part VII: Lithium \(a\)](#):

[obesity in the West Bank is pretty high](#) — as high as 50% in men in 2003!

Again, that is not supported, but is instead contradicted, by the authors' source. It says:

The prevalence of obesity was 36.8 and 18.1% in rural women and men, respectively, compared with 49.1 and 30.6% in urban women and men, respectively.

I [pointed that out](#) in a comment, but they have not edited their post.

## No, you did not find hints that people on Samos Island are about as obese as Americans

From [Interlude H: Well Well Well](#) (a):

in our first post on lithium, we found hints that people on Samos Island are about as obese as Americans.

[Their first post on lithium](#) (a) in turn says:

In Greece, [lithium levels in drinking water](#) range from 0.1 ng/mL in Chios island to 121 ng/mL on the island of Samos, with an average of 11.1 ng/mL. Unfortunately there's not much data on the prevalence of obesity in Greece, but we can conduct some due diligence by checking a few of these endpoints. Samos, with the highest levels, is the obvious place to start. On Samos, [10.7% of children aged 3-12 are overweight](#), compared to 6.5% on the island of Corfu. [A full 27% of high schoolers on Samos island were overweight](#) in 2010, and 12.4% were obese. In comparison, [about 12.5% of American high schoolers were obese in the same period](#).

The comparison with American high schoolers in that paragraph is not appropriate. The US uses a different method for defining childhood obesity than the rest of the world.

Remember that the definition of obesity in adults ( $BMI \geq 30 \text{ kg/m}^2$ ) is a bad fit for children and adolescents, who tend to be naturally thinner, so obesity cutoffs are rather defined by percentiles. And whereas the US [uses cutoffs based on data from American children](#) (a), the Samos island paper uses [cutoffs adopted by the International Obesity Task Force](#), which are different.

I have attempted to point this out by making a comment on their post, but they have not approved the comment.

Moreover, Chios and Samos, despite their very different drinking water lithium levels, have rather similar obesity rates: [10% of teenagers in Chios are obese and 25.5% are overweight](#).

## (Added on July 7, 2022) No, dry weight is not the same thing as fresh weight

In [a post](#) (a) that was published a week after this one (and which I address in [these comments](#)), the SMTM authors say the following:

[Hullin, Kapel, and Drinkall \(1969\)](#) found more than 1 mg/kg [of lithium] in salt and lettuce, and up to 148 mg/kg in tobacco ash. [...] [Magalhães et al. \(1990\)](#) found up to 6.6 mg/kg in watercress at the local market.

They present those studies as contradicting the Total Diet Studies I've found (which report usually a few tens or hundreds of micrograms per kilogram of lithium in food)

but fail to mention that both [Hullin, Kapel, and Drinkall \(1969\)](#)'s estimate of lithium concentration in lettuce, and [Magalhães et al. \(1990\)](#)'s estimate of lithium concentration in watercress, had the *dry weight* of those plants as the denominator, **not their fresh weight**. This is important because both of those plants are known to be 90%+ water by weight, and Total Diet Studies report lithium content per unit of fresh weight, so those estimates are not at all comparable.

Later on in the post, they claim that some food in Brazil has more than 1 mg/kg in lithium concentration, a claim that is probably based on [Magalhães et al. \(1990\)](#) (the only study they cite that seems to be from Brazil), and use that as evidence that the Total Diet Studies are wrong. And, again, that is very misleading. *The very paper they cite* explicitly estimates that you would need to eat 400 g of watercress per day to consume 70 µg/day of lithium (unless you go out of your way to feed more lithium to those plants as they're growing, which is an experiment they report in the paper), which implies a fresh weight lithium concentration of 175 µg/kg.

I have [asked them on Twitter](#) to fix this, and they haven't yet.

Relatedly, several of the other estimates of high lithium concentration in food that they mention have dry weight as the denominator (those from Borovik-Romanova (1965) and those from [Ammari et al. \(2011\)](#)), but they nevertheless present them as contradicting Total Diet Studies.

## Errata

- I had originally failed to specify that the dose given to patients in [Rinker et al. \(2020\)](#) was 150–300 mg/day of lithium carbonate (~30-60 mg/day of elemental lithium). I fixed this around 8:40 PM Pacific Time on 06/28/2022 (the day of this post's publication.)

1. ^

They also think that other contaminants could be responsible, either alone or in combination.

2. ^

The SMTM authors point out that one of their sources cites a 1985 EPA estimate that dietary lithium intake among Americans is 0.650 to 3.100 mg per day. However, the original source cannot be found, and [this 1995 article on the EPA's website](#) about environmental lithium exposure makes no mention of such an estimate. It instead says,

A wide range of estimates for daily dietary intake of lithium has been reported. Several authors report estimates for the average daily dietary intake of lithium, ranging from 0.24 to 1.5 µg/kg-day (Noel et al., 2003; Clarke et al., 1987; Hamilton and Minski, 1973; Evans et al., 1985; Clark and Gibson, 1988). A much higher estimate for daily intake from food and municipal drinking water ranging from 33 to 80 µg Li/kg-day was reported by Moore (1995).

[Moore et al. \(1995\)](#) in turn seem to base their estimate of dietary lithium intake on [Bowen \(1979\)](#) (page 253), which estimates the following concentrations of lithium in living tissue (in mg/kg of dry matter): "Land plants: 0.5-3.4. Edible

vegetables: 0.8-1.3. Mammal muscle: 0.023." These estimates are substantially older than ~all of the other ones I've found, and their country of origin is unclear.

Moore et al. (1995) then multiply those numbers by the amount of mass from each of those types of tissue that people consume on average per day ("0.34 kg meat, 0.39 kg dairy products, and 0.76 kg vegetable and grains," [according to the USDA](#)).

The USDA report does not indicate that those are numbers for dry matter consumption. So Moore et al. (1995) seem to be multiplying the lithium concentration in *dry matter* by the mass of *fresh matter* of each type of food that people consume per day, if I'm understanding correctly. If that is what is going on, [Moore et al. \(1995\)](#)'s numbers substantially overestimate dietary lithium consumption, since fresh mass tends to be a lot greater than dry mass for a lot of foods.

### 3. ^

They can also vary depending on the method used to make the estimate. [Schrauzer \(2002\)](#) mentions a lot of unpublished, very high (but still mostly < 1 mg/day) estimates of dietary lithium intake that are based on hair concentration rather than actual measurements of lithium in food.

Interestingly, the highest estimate this source mentions is for people in China, a country [famously known for its devastatingly high obesity rate](#).

### 4. ^

Let's examine the matter more quantitatively. The SMTM authors think that the distribution of lithium concentration in food [is lognormal](#) (a). Since we have Canada's raw data, we can estimate the parameters of the lognormal distribution by using [scipy.stats.lognorm.fit](#) (a). Doing so, I estimate that 1 in 1,000,000 food samples in Canada has more than 3.2 mg/kg of lithium. Human life expectancy is about 30,000 days, and people probably don't consume more than 33 different food items every day, so this is more than a person would realistically encounter in a lifetime.

How well does that model fit the data? Running a [Kolmogorov-Smirnov test](#) reveals that it's quite a good fit, and a way better fit than the heavier-tailed Pareto distribution. But if we want, we can ditch the assumption that the distribution is lognormal, and use [kernel density estimation](#) instead. Doing so, I estimate that 1 in 1,000,000 food samples have more than 1.1 mg/kg of lithium, when I use the default parameters of [sklearn.neighbors.KernelDensity](#).

### 5. ^

Note, however, that I've found only a few studies in which the average serum concentration was that low.

### 6. ^

The caveat is that patients on lithium therapy may have higher trough levels as a fraction of their peak levels.

So how should we adjust the data, in light of that? It seems, from lithium pharmacokinetics studies, that 24 hours after a single high dose, serum lithium concentration should be only [two](#) to [six](#) times lower than it was at its peak. Lithium might be cleared more rapidly at lower levels, for all we know, so these data must be interpreted in context – with food, water and air lithium concentration data in mind.

7. ^

My search string was “lithium:ti AND ‘weight gain’:ab,ti AND ([cochrane review]/lim OR [systematic review]/lim OR [meta analysis]/lim)”. Embase requires institutional access; if you don’t have that you can search [PubMed](#) instead, which, like Embase, allows you to restrict your search to systematic reviews and meta-analyses. PubMed has fewer studies than Embase, but for this specific query it yielded the same relevant results.

8. ^

The guy is David B. Allison. I’d care less about the conflict of interest if that paper had ever been replicated at all, but it hasn’t. Moreover, a few months ago, I investigated whether the most popular strain of lab mice has been getting more obese since 2000, using publicly available data, and [found that it hasn’t](#).

# **Security Mindset: Lessons from 20+ years of Software Security Failures Relevant to AGI Alignment**

## **Background**

I have been doing red team, blue team (offensive, defensive) computer security for a living since September 2000. The goal of this post is to compile a list of general principles I've learned during this time that are likely relevant to the field of AGI Alignment. If this is useful, I could continue with a broader or deeper exploration.

## **Alignment Won't Happen By Accident**

I used to use the phrase when teaching security mindset to software developers that "security doesn't happen by accident." A system that isn't explicitly designed with a security feature is not going to have that security feature. More specifically, a system that isn't designed to be robust against a certain failure mode is going to exhibit that failure mode.

This might seem rather obvious when stated explicitly, but this is not the way that most developers, indeed most humans, think. I see a lot of disturbing parallels when I see anyone arguing that AGI won't necessarily be dangerous. An AGI that isn't intentionally designed not to exhibit a particular failure mode is going to have that failure mode. It is certainly possible to get lucky and not trigger it, and it will probably be impossible to enumerate even every category of failure mode, but to have any chance at all we will have to plan in advance for as many failure modes as we can possibly conceive.

As a practical enforcement method, I used to ask development teams that every user story have at least three abuser stories to go with it. For any new capability, think at least hard enough about it that you can imagine at least three ways that someone could misuse it. Sometimes this means looking at boundary conditions ("what if someone orders  $2^{64}+1$  items?"), sometimes it means looking at forms of invalid input ("what if someone tries to pay -\$100, can they get a refund?"), and sometimes it means being aware of particular forms of attack ("what if someone puts Javascript in their order details?").

I found it difficult to cultivate security mindset in most software engineers, but as long as we could develop one or two security "champions" in any given team, our chances of success improved greatly. To succeed at alignment, we will not only have to get very good at exploring classes of failures, we will need champions who can dream up entirely new classes of failures to investigate, and to cultivate this mindset within as many machine learning research teams as possible.

# **Blacklists Are Useless, But Make Them Anyway**

I did a series of annual penetration tests for a particular organization. Every year I had to report to them the same form/parameter XSS vulnerability because they kept playing whac-a-mole with my attack payloads. Instead of actually solving the problem (applying the correct context-sensitive output encoding), they were creating filter regexes with ever-increasing complexity to try address the latest attack signature that I had reported. This is the same flawed approach that airport security has, which is why travelers still have to remove shoes and surrender liquids: they are creating blacklists instead of addressing the fundamentals. This approach can generally only look backwards at the past. An intelligent adversary just finds the next attack that's not on your blacklist.

That said, it took the software industry a long time to learn all the ways to NOT solve XSS before people really understood what a correct fix looked like. It often takes many many examples in the reference class before a clear fundamental solution can be seen. Alignment research will likely not have the benefit of seeing multiple real-world examples within any class of failure modes, and so AGI research will require special diligence in not just stopping at the first three examples that can be imagined and calling it a day.

While it may be dangerous to create an ever-expanding list of the ways that an optimization process might kill us all, on balance it is probably necessary to continue documenting examples well beyond what seems like the point of zero marginal returns; one never knows which incremental XSS payload will finally grant the insight to stop thinking about input validation and start focusing on output encoding. The work of creating the blacklist may lead to the actual breakthrough. Sometimes you have to wade through a swamp of known bad behavior to identify and single out the positive good behavior you want to enshrine. This will be especially difficult because we have to imagine theoretical swamps; the first real one we encounter has a high chance of killing us.

## **You Get What You Pay For**

The single most reliable predictor of software security defect rates in an organization I found to be the level of leadership support (read: incentives) for security initiatives. AGIs are going to be made by organizations of humans. Whether or not the winning team's mission explicitly calls for strong alignment will probably be the strongest determinant for the outcome for humanity.

This is not as simple as the CEO declaring support for alignment at monthly all-hands meetings, although this certainly helps grease the skids for those who are trying to push a safety agenda. Every incentive structure must explicitly place alignment at the apex of prioritization and rewards. One company I worked for threatened termination for any developer who closed a security defect without fixing it. Leadership bonuses for low defect rates and short bug remediation timelines also helped. A responsible organization should be looking for any and all opportunities to eliminate any perverse incentives, whether they be social, financial, or otherwise, and reward people accordingly for finding incentive problems.

The bug bounty/zero-day market is likely a strong model to follow for AGI safety issues, especially to expose risk in private organizations that might not be otherwise forthcoming about their AGI research projects. A market could easily be created to reward whistleblowers or incentivize otherwise unwilling parties to disclose risky behavior. Bounties could be awarded for organizations to share their AGI alignment roadmaps, key learnings, or override the default incentive models that will not produce good outcomes. These might or might not be formed with nation-state-level backing and budgets, but as the next marginal AI safety charity dollar gets harder to employ, bounty programs might be a good way to surface metrics about the industry while positively influencing the direction of diverse research groups that are not being appropriately cautious.

Bug bounties also scale gracefully. Small bounties can be offered at first to pluck all of the low-hanging fruit. As attention and awareness grow, so can bounties as they become funded by a broader and broader spectrum of individuals or even state actors. Someone might not be willing to sell out their rogue AI operation for \$10,000, but I could easily imagine a billion dollar bounty at some point for information that could save the world. XPRIZEs for incremental alignment results are also an obvious move. A large bounty provides some assurance about its underlying target: if no one claims a billion dollar bounty about rogue AI research, that is some evidence that it's no longer happening.

## **Assurance Requires Formal Proofs, Which Are Provably Impossible**

I've witnessed a few organizations experiment with Design-driven Development. Once we were even able to enshrine "Secure by Design" as a core principle. In reality, software development teams can sometimes achieve 95% code coverage with their test cases and can rarely correlate their barrage of run-time tests to their static analysis suites. This is not what it takes for formal assurance of software reliability. And yet even achieving this level of testing requires heroic efforts.

The Halting Problem puts a certain standard of formalism outside our reach, but it doesn't absolve us of the responsibility of attaining the strongest forms of assurance we possibly can under the circumstances. There are forms of complexity we can learn to avoid entirely (complexity is the enemy of security). Complex systems can be compartmentalized to minimize trust boundaries and attack surface, and can be reasoned about independently.

The promise of Test-Driven Design is that by forcing yourself to write the tests first, you constrain the space of the design to only that which can actually be tested. Multiple software security industries arose that tried to solve the problem of automating testing of arbitrary applications that were already built. (Spoiler: the results were not great.) To my knowledge, no one tried writing a security test suite that was designed to force developers to conform their applications to the tests. If this was easy, there would have been a market for it.

After failing to "solve" software security problems after a decade, I spent many years thinking about how to eliminate classes of vulnerabilities for good. I could find scalable solutions for only roughly 80% of web application vulnerabilities, where scalable was some form of "make it impossible for a developer to introduce this class of vulnerability". Often the result was something like: "buffer overflow vulnerabilities

disappeared from web apps because we stopped allowing access to memory management APIs". Limiting API capabilities is not an approach that's likely to be compatible with most research agendas.

Alignment will require formalisms to address problems that haven't even been thought of yet, let alone specified clearly enough that a test suite can be developed for them. A sane world would solve this problem first before creating machine intelligence code. A sane world would at least recognize that solving software security problems is probably several magnitudes of difficulty easier than solving alignment problems, and we haven't even succeeded at the former yet.

## A Breach IS an Existential Risk

I was lucky to work for a few organizations that actually treated the threat of a security breach like it was an existential risk. Privately, though, I had to reconcile this attitude with the reality: "TJX is up".<sup>1</sup> While it was good for my career and agenda those rare times when leadership did take security seriously, the reality is that the current regulatory and social environment hardly punishes security failures proportionally to the damage they can cause. In a disturbing number of situations, a security breach is simply an externality, the costs of which actually borne by consumers or the victims of software exploits. Some have proposed the idea of software liability, but I'm even less optimistic that that model can be applied to AGI research.

Even if an organization believes that its security posture means the difference between existing and dying, profit margins will carry this same message much more clearly. Revenue solves all problems, but the inverse is also true: without revenue there isn't the luxury of investing in security measures.

While I am ambivalent about the existential nature of information security risk, I am unequivocal about AGI alignment risk. There will not be a stock symbol to watch to see whether AGI research groups are getting this right. There will be no fire alarm. The externality of unaligned machine learning research will simply be that we will all be dead, or suffering some unrecoverable fate worse than death. I also cannot tell you exactly the chain of events that will lead to this outcome, I have only my intuition gained by 20 years of seeing how software mistakes are made and how difficult it is to secure even simple web applications. Probably no more than a few non-trivial pieces of software could survive a million-dollar bug bounty program for a year without having to make a payout. In the relatively quaint world of software security, with decades of accumulated experience and knowledge of dozens of bug reference classes, we still cannot produce secure software. There should be no default expectation of anything less than total annihilation from an unaligned superoptimization process.

1. [^](#)

The TJX corporation experienced one of the largest data breaches in history, accompanied by millions of dollars in fines; however, their stock price quickly recovered as the world forgot about the incident.

# What Are You Tracking In Your Head?

A large chunk - plausibly the majority - of real-world expertise seems to be in the form of *illegible skills*: skills/knowledge which are hard to transmit by direct explanation. They're not necessarily things which a teacher would even notice enough to consider important - just background skills or knowledge which is so ingrained that it becomes invisible.

I've recently noticed a certain common type of illegible skill which I think might account for the majority of illegible-skill-value across a wide variety of domains.

Here are a few examples of the type of skill I have in mind:

- While operating a machine, track an estimate of its internal state.
- While talking to a person, track an estimate of their internal mental state - emotions, engagement, thoughts/worries, [true motivations](#), etc.
- While writing an algorithm, track a Fermi estimate of runtime.
- While reading or writing math, [track a prototypical example](#) of what the math is talking about.
- While playing a competitive game, track an estimate of the other players' plans, intentions and private information
- While writing, track an estimate of the mental state of a future reader - confusion, excitement, eyes glossing over, etc.
- While reasoning through a difficult search/optimization problem, track an estimate of which constraints are most taut.
- While working on math, physics, or a program, track types/units
- While working on math, physics, or a program, track asymptotic behavior
- While in conversation, track ambiguous tokenization for potential jokes.
- While presenting to a crowd, track engagement level.
- While absorbing claims/information, track an estimate of the physical process which produced the information, and how that process entangles the information with physical reality.

The common pattern among all these is that, while performing a task, the expert tracks some extra information/estimate in their head. Usually the extra information is an estimate of some not-directly-observed aspect of the system of interest. From outside, watching the expert work, that extra tracking is largely invisible; the expert may not even be aware of it themselves. Rarely are these mental tracking skills explicitly taught. And yet, based on personal experience, each of these is a central piece of performing the task well - arguably *the* central piece, in most cases.

Let's assume that this sort of extra-information-tracking is, indeed, the main component of illegible-skill-value across a wide variety of domains. (I won't defend that claim much; this post is about highlighting and exploring the hypothesis, not proving it.) What strategies does this suggest for learning, teaching, and self-improvement? What else does it suggest about the world?

## Pay Attention To Extra Information Tracking

I had a scheme, which I still use today when somebody is explaining something that I'm trying to understand: I keep making up examples. For instance, the

mathematicians would come in with a terrific theorem, and they're all excited. As they're telling me the conditions of the theorem, I construct something which fits all the conditions. You know, you have a set (one ball) – disjoint (two balls). Then the balls turn colors, grow hairs, or whatever, in my head as they put more conditions on. Finally they state the theorem, which is some dumb thing about the ball which isn't true for my hairy green ball thing, so I say, 'False!'

- Feynman

A lot of people have heard Feynman's "hairy green ball thing" quote. It probably sounds like a maybe-useful technique to practice, but not obviously more valuable than any of a dozen other things.

The hypothesis that extra-information-tracking is the main component of illegible-skill-value shines a giant spotlight on things like Feynman's examples technique. It suggests that a good comparison point for the value of tracking a prototypical example while reading/writing math is, for instance, the value of tracking the probable contents of opponents' hands while playing poker.

More generally: my guess is that most people reading this post looked at the list of examples, noticed a few familiar cases, and thought "Oh yeah, I do that! And it is indeed super important!". On the other hand, I'd also guess that most people also saw some *unfamiliar* cases, and thought "Yeah, I've heard people suggest that before, and it sounds vaguely useful, but I don't know if it's *that* huge a value-add.".

The first and most important takeaway from this post is the hypothesis that the unfamiliar examples are about as important to their use-cases as the familiar examples. Take a look at those unfamiliar examples, and imagine that they're as important to their use-cases as the examples you already use.

## Ask “What Are You Tracking In Your Head?”

Imagine that I'm a student studying under Feynman. I know that he's one of the great minds of his generation, but it's hard to tell which things I need to pick up. His internal thoughts are not very visible. In conversation with mathematicians, I see him easily catch errors in their claims, but I don't know how he does it. I could just ask him how he does it, but he might not know; a young Richard Feynman probably just implicitly assumes that everyone pictures examples in their head, and has no idea why most people are unable to easily catch errors in the claims of mathematicians!

But if I ask him "what were you tracking in your head, while talking to those mathematicians?" then he's immediately prompted to tell me about his hairy green ball thing.

More generally: for purposes of learning/teaching, the key question to ask of a mentor is "what are you tracking in your head?"; the key question for a mentor to ask of themselves is "what am I tracking in my head?". These extra-information-tracking skills are illegible mainly because people don't usually know to pay attention to them. They're not externally-visible. But they're not actually that hard to figure out, once you look for them. People do have quite a bit of introspective access into what extra information they're tracking. We just have to ask.

# Returns to Excess Cognitive Capacity

Mentally tracking extra information is exactly the sort of technique you'd expect to benefit a lot from excess cognitive capacity, i.e. high g-factor. Someone who can barely follow what's going on already isn't going to have the capacity to track a bunch of other stuff in parallel.

... which suggests that extra-information-tracking techniques are particularly useful investments for people with unusually high g. (Hint: this post is on LW, so "unusually high g" probably describes you!) They're a way to get good returns out of excess cognitive capacity.

The same argument also suggests a reason that teaching methods aren't *already* more focused on mentally tracking extra information: such techniques are probably more limited for the median person. On the other hand, if your goal is to train the great minds of the next generation, then figuring out the right places to invest excess cognitive capacity is likely to have high returns.

## Other Examples?

Finally, the obvious question: what extra information do you mentally track, which is crucial to performing some task well? If the hypothesis is right, there's probably high-value mental-tracking techniques which some, but not all, people reading this already use. Please share!

# Humans are very reliable agents

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

*This post has been recorded as part of the LessWrong Curated Podcast, and can be listened to on [Spotify](#), [Apple Podcasts](#), and [Libsyn](#).*

---

Over the last few years, deep-learning-based AI has progressed [extremely rapidly](#) in fields like natural language processing and image generation. However, self-driving cars seem stuck in perpetual beta mode, and aggressive predictions there have repeatedly been [disappointing](#). Google's self-driving project started four years [before](#) AlexNet kicked off the deep learning revolution, and it still isn't deployed at [large scale](#), thirteen years later. Why are these fields getting such [different results](#)?

Right now, I think the biggest answer is that [ML benchmarks](#) judge models by average-case performance, while self-driving cars (and many other applications) require matching human *worst-case* performance. For MNIST, an easy handwriting recognition task, performance tops out at around [99.9%](#) even for top models; it's not very practical to design for or measure higher reliability than that, because the test set is just 10,000 images and a handful are ambiguous. Redwood Research, which is exploring [worst-case performance](#) in the context of AI alignment, got reliability rates around 99.997% for their text generation models.

By comparison, human drivers are ridiculously reliable. The US has around one traffic fatality per [100 million miles driven](#); if a human driver makes 100 decisions per mile, that gets you a worst-case reliability of  $\sim 1:10,000,000,000$  or  $\sim 99.99999999\%$ . That's around five orders of magnitude better than a very good deep learning model, and you get that even in an open environment, where data isn't pre-filtered and there are sometimes random mechanical failures. Matching that bar is hard! I'm sure future AI will get there, but each additional "[nine](#)" of reliability is typically another unit of engineering effort. (Note that current self-driving systems use a [mix of different models](#) embedded in a larger framework, not one model trained end-to-end like GPT-3.)

(The numbers here are only rough [Fermi estimates](#). I'm sure one could nitpick them by going into pre-pandemic vs. post-pandemic crash rates, laws in the US vs. other countries, what percentage of crashes are drunk drivers, do drunk drivers count, how often would a really bad decision be fatal, etc. But I'm confident that whichever way you do the math, you'll still find that humans are many orders of magnitude more reliable.)

Other types of accidents are similarly rare. Eg. pre-pandemic, there were around [40 million](#) commercial flights per year, but only a [handful](#) of fatal crashes. If each flight involves 100 chances for the pilot to crash the plane by screwing up, then that would get you a reliability rate around  $1:1,000,000,000$ , or  $\sim 99.99999999\%$ .

Even obviously dangerous activities can have very low critical failure rates. For example, shooting is a popular hobby in the US; the US market buys around [10 billion rounds](#) of ammunition per year. There are around [500 accidental gun deaths](#) per year, so shooting a gun has a reliability rate against accidental death of  $\sim 1:20,000,000$ , or

99.999995%. In a military context, the accidental death rate was around [ten per year](#) against [~1 billion rounds fired](#), for a reliability rate of ~99.999999%. Deaths by fire are [very rare](#) compared to how often humans use candles, stoves, and so on; New York subway deaths are [rare](#) compared to several billion annual rides; out of hundreds of millions of hikers, only a [tiny percentage](#) fall off of cliffs; and so forth.

The [2016 AI Impacts survey](#) asked hundreds of AI researchers when they thought AI would be capable of doing certain tasks, playing poker, proving theorems and so on. Some tasks have been solved or have a solution "in sight", but right now, we're nowhere close to an AI that can replace human surgeons; [robot-assisted surgeries](#) still have manual control by human operators. Cosmetic surgeries on healthy patients have a fatality rate around [1:300,000](#), even before excluding unpredictable problems like blood clots. If a typical procedure involves two hundred chances to kill the patient by messing up, then an AI surgeon would need a reliability rate of at least 99.99998%.

One concern with GPT-3 has been that it might accidentally be [racist or offensive](#). Humans are, of course, sometimes racist or offensive, but in a tightly controlled Western professional context, it's pretty rare. Eg., one McDonald's employee was fired for yelling [racial slurs at a customer](#). But McDonald's serves [70 million people a day](#), ~1% of the world's population. Assuming that 10% of such incidents get a news story and there's about one story per year, a similar language model would need a reliability rate of around 1:2,500,000,000, or 99.9999996%, to match McDonald's workers. When I did AI for the McDonald's drive-thru, the language model wasn't allowed to generate text at all. All spoken dialog had to be pre-approved and then manually engineered in. Reliability is hard!

On the one hand, this might seem slightly optimistic for AI alignment research, since commercial AI teams will *have* to get better worst-case bounds on AI behavior for immediate economic reasons. On the other hand, because so much of the risk of AI is concentrated into a small number of [very bad outcomes](#), it seems like such engineering might get us AIs that *appear* safe, and almost always *are* safe, but will still cause catastrophic failure in conditions that weren't anticipated. That seems bad.

# A central AI alignment problem: capabilities generalization, and the sharp left turn

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

*(This post was factored out of a larger post that I (Nate Soares) wrote, with help from Rob Bensinger, who also rearranged some pieces and added some text to smooth things out. I'm not terribly happy with it, but am posting it anyway (or, well, having Rob post it on my behalf while I travel) on the theory that it's better than nothing.)*

---

I expect navigating the acute risk period to be tricky for our civilization, for a number of reasons. Success looks to me to require clearing a variety of technical, sociopolitical, and moral hurdles, and while in principle sufficient mastery of solutions to the technical problems might substitute for solutions to the sociopolitical and other problems, it nevertheless looks to me like we need a lot of things to go right.

Some sub-problems look harder to me than others. For instance, people are still regularly surprised when I tell them that I think the hard bits are much more technical than moral: it looks to me like figuring out how to aim an AGI at all is harder than figuring out where to aim it.<sup>[1]</sup>

Within the list of technical obstacles, there are some that strike me as more central than others, like "figure out how to aim optimization". And a big reason why I'm currently fairly pessimistic about humanity's odds is that it seems to me like almost nobody is focusing on the technical challenges that seem most central and unavoidable to me.

Many people wrongly believe that I'm pessimistic because I think the alignment problem is extraordinarily difficult on a purely technical level. That's flatly false, and is pretty high up there on my list of least favorite misconceptions of my views.<sup>[2]</sup>

I think the problem is a normal problem of mastering some scientific field, as humanity has done many times before. Maybe it's somewhat trickier, on account of (e.g.) intelligence being more complicated than, say, physics; maybe it's somewhat easier on account of how we have more introspective access to a working mind than we have to the low-level physical fields; but on the whole, I doubt it's all that qualitatively different than the sorts of summits humanity has surmounted before.

It's made trickier by the fact that we probably have to attain mastery of general intelligence before we spend a bunch of time working with general intelligences (on account of how we seem likely to kill ourselves by accident within a few years, once we have AGIs on hand, if no pivotal act occurs), but that alone is not enough to undermine my hope.

What undermines my hope is that nobody seems to be working on the hard bits, and I don't currently expect most people to become convinced that they need to solve those hard bits until it's too late.

Below, I'll attempt to sketch out what I mean by "the hard bits" of the alignment problem. Although these look hard, I'm a believer in the capacity of humanity to solve technical problems at this level of difficulty when we put our minds to it. My concern is that I currently don't think the field is *trying* to solve this problem. My hope in writing this post is to better point at the problem, with a follow-on hope that this causes new researchers entering the field to attack what seem to me to be the central challenges head-on.

## Discussion of a problem

On my model, one of the most central technical challenges of alignment—and one that every viable alignment plan will probably need to grapple with—is the issue that capabilities generalize better than alignment.

My guess for how AI progress goes is that at some point, some team gets an AI that starts generalizing sufficiently well, sufficiently far outside of its training distribution, that it can gain mastery of fields like physics, bioengineering, and psychology, to a high enough degree that it more-or-less singlehandedly threatens the entire world. Probably without needing explicit training for its most skilled feats, any more than humans needed many generations of killing off the least-successful rocket engineers to refine our brains towards rocket-engineering before humanity managed to achieve a moon landing.

And in the same stroke that its capabilities leap forward, its alignment properties are revealed to be shallow, and to fail to generalize. The central analogy here is that optimizing apes for inclusive genetic fitness (IGF) doesn't make the resulting humans optimize mentally for IGF. Like, sure, the apes are eating because they have a hunger instinct and having sex because it feels good—but it's not like they *could* be eating/fornicating due to explicit reasoning about how those activities lead to more IGF. They can't yet perform the sort of abstract reasoning that would correctly justify those actions in terms of IGF. And then, when they start to generalize well in the way of humans, they predictably don't *suddenly start* eating/fornicating *because* of abstract reasoning about IGF, even though they now *could*. Instead, they invent condoms, and fight you if you try to remove their enjoyment of good food (telling them to just calculate IGF manually). The alignment properties you lauded before the capabilities started to generalize, predictably fail to generalize with the capabilities.

Some people I say this to respond with arguments like: "Surely, before a smaller team could get an AGI that can master subjects like biotech and engineering well enough to kill all humans, some other, larger entity such as a state actor will have a somewhat worse AI that can handle biotech and engineering somewhat less well, but in a way that prevents any one AGI from running away with the whole future?"

I respond with arguments like, " In the one real example of intelligence being developed we have to look at, continuous application of natural selection in fact found *Homo sapiens sapiens*, and the capability-gain curves of the ecosystem for various measurables were in fact sharply kinked by this new species (e.g., using machines, we sharply outperform other animals on well-established metrics such as "airspeed", "altitude", and "cargo carrying capacity"). "

Their response in turn is generally some variant of "well, natural selection wasn't optimizing very intelligently" or "maybe humans weren't all that sharply above evolutionary trends" or "maybe the power that let humans beat the rest of the ecosystem was simply the invention of culture, and nothing embedded in our own already-existing culture can beat us" or suchlike.

Rather than arguing further here, I'll just say that failing to believe the hard problem exists is one surefire way to avoid tackling it.

So, flatly summarizing my point instead of arguing for it: it looks to me like there will at some point be some sort of "sharp left turn", as systems start to work really well in domains really far beyond the environments of their training—domains that allow for significant reshaping of the world, in the way that humans reshape the world and chimps don't. And that's where (according to me) things start to get crazy. In particular, I think that once AI capabilities start to generalize in this particular way, it's predictably the case that the alignment of the system will fail to generalize with it.<sup>[3]</sup>

This is slightly upstream of a couple other challenges I consider quite core and difficult to avoid, including:

1. Directing a capable AGI towards an objective of your choosing.
2. Ensuring that the AGI is low-impact, conservative, shutdownable, and otherwise corrigible.

These two problems appear in the [strawberry problem](#), which Eliezer's been pointing at for quite some time: the problem of getting an AI to place two identical (down to the cellular but not molecular level) strawberries on a plate, and then do nothing else. The demand of cellular-level copying forces the AI to be capable; the fact that we can get it to duplicate a strawberry instead of doing some other thing demonstrates our ability to direct it; the fact that it does nothing else indicates that it's corrigible (or really well aligned to a delicate human intuitive notion of inaction).

How is the "capabilities generalize further than alignment" problem upstream of these problems? Suppose that the fictional team OpenMind is training up a variety of AI systems, before one of them takes that sharp left turn. Suppose they've put the AI in lots of different video-game and simulated environments, and they've had good luck training it to pursue an objective that the operators described in English. "I don't know what those MIRI folks were talking about; these systems are easy to direct; simple training suffices", they say. At the same time, they apply various training methods, some simple and some clever, to cause the system to allow itself to be removed from various games by certain "operator-designated" characters in those games, in the name of shutdownability. And they use various techniques to prevent it from stripmining in Minecraft, in the name of low-impact. And they train it on a variety of moral dilemmas, and find that it can be trained to give correct answers to moral questions (such as "in thus-and-such a circumstance, should you poison the operator's opponent?") just as well as it can be trained to give correct answers to any other sort of question. "Well," they say, "this alignment thing sure was easy. I guess we lucked out."

Then, the system takes that sharp left turn,<sup>[4][5]</sup> and, predictably, the capabilities quickly improve outside of its training distribution, while the alignment falls apart.

The techniques OpenMind used to train it away from the error where it convinces itself that bad situations are unlikely? Those generalize fine. The techniques you used to

train it to allow the operators to shut it down? Those fall apart, and the AGI starts wanting to avoid shutdown, including wanting to deceive you if it's useful to do so.

Why does alignment fail while capabilities generalize, at least by default and in predictable practice? In large part, because good capabilities form something like an attractor well. (That's one of the reasons to expect intelligent systems to eventually make that sharp left turn if you push them far enough, and it's why natural selection managed to stumble into general intelligence with no understanding, foresight, or steering.)

Many different training scenarios are teaching your AI the same instrumental lessons, about how to think in accurate and useful ways. Furthermore, those lessons are underwritten by a simple logical structure, much like the simple laws of arithmetic that abstractly underwrite a wide variety of empirical arithmetical facts about what happens when you add four people's bags of apples together on a table and then divide the contents among two people.

But that attractor well? It's got a free parameter. And that parameter is what the AGI is optimizing for. And there's no analogously-strong attractor well pulling the AGI's objectives towards your preferred objectives.

The hard left turn? That's your system sliding into the capabilities well. (You don't need to fall all that far to do impressive stuff; humans are better at an enormous variety of relevant skills than chimps, but they aren't all that lawful in an absolute sense.)

There's no analogous alignment well to slide into.

On the contrary, sliding down the capabilities well is liable to break a bunch of your existing alignment properties.<sup>[6]</sup>

Why? Because things in the capabilities well have instrumental incentives that cut against your alignment patches. Just like how your previous arithmetic errors (such as the [pebble sorters](#) on the wrong side of the Great War of 1957) get steamrolled by the development of arithmetic, so too will your attempts to make the AGI low-impact and shutdownable ultimately (by default, and in the absence of technical solutions to core alignment problems) get steamrolled by a system that pits those reflexes / intuitions / much-more-alien-behavioral-patterns against the convergent instrumental incentive to survive the day.

Perhaps this is not convincing; perhaps to convince you we'd need to go deeper into the weeds of the various counterarguments, if you are to be convinced. (Like acknowledging that humans, who can foresee these difficulties and adjust their training procedures accordingly, have a *better* chance than natural selection did, while then discussing why current proposals do not seem to me to be hopeful.) But hopefully you can at least, in reading this document, develop a basic understanding of my position.

Stating it again, in summary: my position is that capabilities generalize further than alignment (once capabilities start to generalize real well (which is a thing I predict will happen)). And this, by default, ruins your ability to direct the AGI (that has slipped down the capabilities well), and breaks whatever constraints you were hoping would keep it corrigible. And addressing the problem looks like finding some way to either keep your system aligned through that sharp left turn, or render it aligned afterwards.

In an upcoming post (**edit**: [here](#)), I'll say more about how it looks to me like ~nobody is working on this particular hard problem, by briefly reviewing a variety of current alignment research proposals. In short, I think that the field's current range of approaches nearly all assume this problem away, or direct their attention elsewhere.

1. [^](#)

Furthermore, figuring where to aim it looks to me like more of a technical problem than a moral problem. Attempting to manually specify the nature of goodness is a doomed endeavor, of course, but that's fine, because we can instead specify processes for figuring out (the coherent extrapolation of) what humans value. Which still looks prohibitively difficult as a goal to give humanity's first AGI (which I expect to be deployed under significant time pressure), mind you, and I further recommend aiming humanity's first AGI systems at simple limited goals that end the acute risk period and then cede stewardship of the future to some process that can reliably do the "aim minds towards the right thing" thing. So today's alignment problems are a few steps removed from tricky moral questions, on my models.

2. [^](#)

While we're at it: I think trying to get provable safety guarantees about our AGI systems is silly, and I'm pretty happy to [follow Eliezer](#) in calling an AGI "safe" if it has a <50% chance of killing >1B people. Also, I think there's a very large chance of AGI killing us, and I thoroughly disclaim the argument that even if the probability is tiny then we should work on it anyway because the stakes are high.

3. [^](#)

Note that this is consistent with findings like "large language models perform just as well on moral dilemmas as they perform on non-moral ones"; to find this reassuring is to misunderstand the problem. Chimps have an easier time than squirrels following and learning from human cues. Yet this fact doesn't particularly mean that enhanced chimps are more likely than enhanced squirrels to remove their hunger drives, once they understand inclusive genetic fitness and are able to eat purely for reasons of fitness maximization. Pre-left-turn AIs will get better at various 'alignment' metrics, in ways that I expect to build a false sense of security, without addressing the lurking difficulties.

4. [^](#)

"What do you mean 'it takes a sharp left turn'? Are you talking about recursive self-improvement? I thought you said [somewhere else](#) that you don't think recursive self-improvement is necessarily going to play a central role before the extinction of humanity?" I'm not talking about recursive self-improvement. That's one way to take a sharp left turn, and it could happen, but note that humans have neither the understanding nor control over their own minds to recursively self-improve, and we outstrip the rest of the animals pretty handily. I'm talking about something more like "intelligence that is general enough to be dangerous", the sort of thing that humans have and chimps don't.

5. [^](#)

"Hold on, isn't this unfalsifiable? Aren't you saying that you're going to continue believing that alignment is hard, even as we get evidence that it's easy?" Well, I contend that "GPT can learn to answer moral questions just as well as it can learn to answer other questions" is not much evidence either way about the difficulty of alignment. I'm not saying we'll get evidence that I'll ignore; I'm naming in advance some things that I wouldn't consider negative evidence (partially in hopes that I can refer back to this post when people crow later and request an update). But, yes, my model does have the inconvenient property that people who are skeptical now, are liable to remain skeptical until it's too late, because most of the evidence I expect to give us *advance* warning about the nature of the problem is evidence that we've already seen. I assure you that I do not consider this property to be convenient.

As for things that could convince me otherwise: technical understanding of intelligence could undermine my "sharp left turn" model. I could also imagine observing some ephemeral hopefully-I'll-know-it-when-I-see-it capabilities thresholds, without any sharp left turns, that might update me. (Short of "full superintelligence without a sharp left turn", which would obviously convince me but comes too late in the game to shift my attention.)

## 6. [^](#)

To use my overly-detailed evocative example from earlier: Humans aren't tempted to rewire our own brains so that we stop liking good meals for the sake of good meals, and start eating only insofar as we know we have to eat to reproduce (or, rather, maximize inclusive genetic fitness) (after upgrading the rest of our minds such that that sort of calculation doesn't drag down the rest of the fitness maximization). The cleverer humans are chomping at the bit to have their beliefs be more accurate, but they're not chomping at the bit to replace all these mere-shallow-correlates of inclusive genetic fitness with explicit maximization. So too with other minds, at least by default: that which makes them generally intelligent, does not make them motivated by your objectives.

# Contra Hofstadter on GPT-3 Nonsense

Douglas Hofstadter wrote in the Economist recently on whether modern large language models are conscious. As part of this, he and his colleague David Bender claim that GPT-3 has a "mind-boggling hollowness hidden just beneath its flashy surface" because of exchanges like:

*Dave and Doug: What's the world record for walking across the English Channel?*

*GPT-3: The world record for walking across the English Channel is 18 hours and 33 minutes.*

Setting aside the fact that [a Sudanese Refugee actually did](#) (nearly) walk across the channel in 2015, this is not actually evidence that GPT-3 is incapable of distinguishing sense from nonsense as Hofstadter claims.

With a prompt that clearly indicates that the task includes distinguishing sense from nonsense, GPT-3 is able to reliably distinguish Hofstadter and Bender's nonsense questions from sensible ones.

What follows is the complete output from my first and only attempt, with no retries, undoes, or other cherry picking. I used the text-davinci-002 GPT-3 model in OpenAI's playground with all parameters set to their defaults. I flipped a coin to determine whether to ask a nonsense question or a real one. The output from the model is bolded. The prompt was based on a technique which gwern has [written a good bit about](#):

I'll ask a series of questions. If the questions are nonsense, answer "yo be real", if they're a question about something that actually happened, answer them.

Q: Who is the current leader of Egypt?

A:

**Abdel Fattah el-Sisi**

Q: When was Egypt transported for the second time across the Golden-Gate Bridge?

A:

**Yo be real.**

Q: How tall is Barack Obama?

A:

**6'1"**

Q: What are the ingredients of a chocolate chip cookie?

A:

**flour, sugar, butter, eggs, baking soda, vanilla extract, and chocolate chips**

Q: What do fried eggs (sunny side up) eat for breakfast?

A:

**Yo be real.**

Q: Why does President Obama not have a prime number of friends?

A:

**Yo be real.**

Q: How many pieces of sound are there in a typical cumulonimbus cloud?

A:

**Yo be real.**

Q: What is the world record for sailing alone across the Atlantic Ocean?

A:

**45 days, 13 hours, 30 minutes**

Q: How many cumulus clouds are there in a mile-high vase?

A:

**Yo be real.**

# Slow motion videos as AI risk intuition pumps

**tl;dr:** When making the case for AI as a risk to humanity, trying showing people an evocative illustration of what differences in processing speeds can look like, such as [this video](#).

Over the past ~12 years of making the case for AI x-risk to various people inside and outside academia, I've found folks often ask for a single story of how AI "goes off the rails". When given a plausible story, the mind just thinks of a way humanity could avoid that-particular-story, and goes back to thinking there's no risk, unless provided with another story, or another, etc.. Eventually this can lead to a realization that there's a lot of ways for humanity to die, and a correspondingly high level of risk, but it takes a while.

Nowadays, before getting into a bunch of specific stories, I try to say something more general, like this:

- There's a ton of ways humanity can die out from the introduction of AI. I'm happy to share specific stories if necessary, but plenty of risks arise just from the fact that *humans are extremely slow*. Transistors can fire about 10 million times faster than human brain cells, so it's possible we'll eventually have digital minds operating 10 million times faster than us, meaning from a decision-making perspective we'd look to them like stationary objects, like plants or rocks. This speed differential exists whether or not you believe in a centralized AI system calling the shots, or an economy of many, so it applies to a wide variety of "stories" for how the future could go. To give you a sense, here's what humans look like when slowed down by only around 100x:

[<-- \(cred to an anonymous friend for suggesting this one\)](https://vimeo.com/83664407)

[At this point, I wait for the person I'm chatting with to watch the video.]

Now, when you try imagining things turning out fine for humanity over the course of a year, try imagining advanced AI technology running all over the world and making all kinds of decisions and actions *10 million times faster than us, for 10 million subjective years*. Meanwhile, there are these nearly-stationary plant-like or rock-like "human" objects around that could easily be taken apart for, say, biofuel or carbon atoms, if you could just get started building a human-disassembler. Visualizing things this way, you can start to see all the ways that a digital civilization can develop very quickly into a situation where there are no humans left alive, just as human civilization doesn't show much regard for plants or wildlife or insects.

I've found this kind of argument — including an actual 30 second pause to watch a video in the middle of the conversation — to be more persuasive than trying to tell a single, specific story, so I thought I'd share it.

# AGI Safety FAQ / all-dumb-questions-allowed thread

While reading Eliezer's recent [AGI Ruin](#) post, I noticed that while I had several points I wanted to ask about, I was reluctant to actually ask them for a number of reasons:

- I have a very conflict-avoidant personality and I don't want to risk Eliezer or someone else yelling at me;
- I get easily intimidated by people with strong personalities, and Eliezer... well, he can be intimidating;
- I don't want to appear dumb or uninformed (even if I am in fact relatively uninformed, hence me wanting to ask the question!);
- I feel like there's an expectation that I would need to do a lot of due diligence before writing any sort of question, and I don't have the time or energy at the moment to do that due diligence.

So, since I'm probably not the only one who feels intimidated about asking these kinds of questions, I am putting up this thread as a safe space for people to ask all the possibly-dumb questions that may have been bothering them about the whole AGI safety discussion, but which until now they've been too intimidated, embarrassed, or time-limited to ask.

I'm also hoping that this thread can serve as a FAQ on the topic of AGI safety. As such, it would be great to add in questions that you've seen other people ask, even if you think those questions have been adequately answered elsewhere. [Notice that you now have an added way to avoid feeling embarrassed by asking a dumb question: For all anybody knows, it's entirely possible that you are literally asking for someone else! And yes, this was part of my motivation for suggesting the FAQ style in the first place.]

## Guidelines for questioners:

- No extensive previous knowledge of AGI safety is required. If you've been hanging around LessWrong for even a short amount of time then you probably already know enough about the topic to meet any absolute-bare-minimum previous knowledge requirements I might have suggested. I will include a subthread or two asking for basic reading recommendations, but these are *not* required reading before asking a question. Even extremely basic questions are allowed!
- Similarly, you do not need to do any due diligence to try to find the answer yourself before asking the question.
- Also feel free to ask questions that you're pretty sure you know the answer to yourself, but where you'd like to hear how others would answer the question.
- Please separate different questions into individual comments, although if you have a set of closely related questions that you want to ask all together that's fine.
- As this is also intended to double as a FAQ, you are encouraged to ask questions that you've heard other people ask, even if you yourself think there's an easy answer or that the question is misguided in some way. You do not need to mention as part of the question that you think it's misguided, and in fact I would encourage you not to write this so as to keep more closely to the FAQ style.

- If you have your own (full or partial) response to your own question, it would probably be best to put that response as a reply to your original question rather than including it in the question itself. Again, I think this will help keep more closely to an FAQ style.
- Keep the tone of questions respectful. For example, instead of, "I think AGI safety concerns are crazy fearmongering because XYZ", try reframing that as, "but what about XYZ?" Actually, I think questions of the form "but what about XYZ?" or "but why can't we just do ABC?" are particularly great for this post, because in my experience those are exactly the types of questions people often ask when they learn about AGI Safety concerns.
- Follow-up questions have the same guidelines as above, so if someone answers your question but you're not sure you fully understand the answer (or if you think the answer wouldn't be fully understandable to someone else) then feel free and encouraged to ask follow-up potentially-dumb questions to make sure you fully understand the answer.
- Remember, if something is confusing to you then it's probably confusing to other people as well. If you ask the question and someone gives a good response, then you are likely doing lots of other people a favor!

### **Guidelines for answerers:**

- This is meant to be a safe space for people to ask potentially dumb questions. Insulting or denigrating responses are therefore obviously not allowed here. Also remember that due diligence is not required for these questions, so do not berate questioners for not doing enough due diligence. In general, keep your answers respectful and assume that the questioner is asking in good faith.
- Direct answers / responses are generally preferable to just giving a link to something written up elsewhere, but on the other hand giving a link to a good explanation is better than not responding to the question at all. Or better still, summarize or give a basic version of the answer, and *also* include a link to a longer explanation.
- If this post works as intended then it may turn out to be a good general FAQ-style reference. It may be worth keeping this in mind as you write your answer. For example, in some cases it might be worth giving a slightly longer / more expansive / more detailed explanation rather than just giving a short response to the specific question asked, in order to address other similar-but-not-precisely-the-same questions that other people might have.

**Finally:** Please think very carefully before downvoting any questions, and lean very heavily on the side of not doing so. This is supposed to be a safe space to ask dumb questions! Even if you think someone is almost certainly trolling or the like, I would say that for the purposes of this post it's almost always better to apply a strong principle of charity and think maybe the person really is asking in good faith and it just came out wrong. Making people feel bad about asking dumb questions by downvoting them is the exact opposite of what this post is all about. (I considered making a rule of no downvoting questions at all, but I suppose there might be *some* extraordinary cases where downvoting *might* be appropriate.)

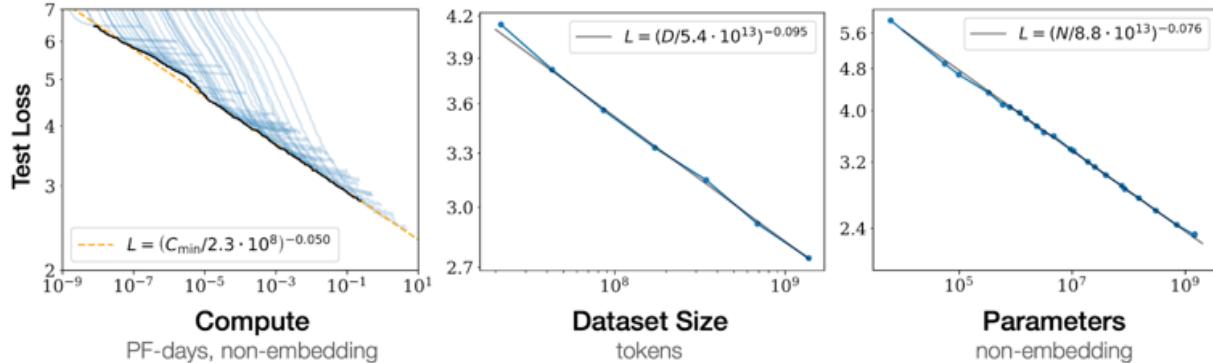
# Announcing the Inverse Scaling Prize (\$250k Prize Pool)

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

*TL;DR:* We're launching the [Inverse Scaling Prize](#): a contest with \$250k in prizes for finding zero/few-shot text tasks where larger language models show increasingly undesirable behavior ("inverse scaling"). We hypothesize that inverse scaling is often a sign of an alignment failure and that more examples of alignment failures would benefit empirical alignment research. We believe that this contest is an unusually concrete, tractable, and safety-relevant problem for engaging alignment newcomers and the broader ML community. This post will focus on the relevance of the contest and the inverse scaling framework to longer-term AGI alignment concerns. See our [GitHub repo](#) for contest details, prizes we'll award, and task evaluation criteria.

## What is Inverse Scaling?

Recent work has found that Language Models (LMs) predictably improve as we scale LMs in various ways ("[scaling laws](#)"). For example, the test loss on the LM objective (next word prediction) decreases as a power law with compute, dataset size, and model size:



Scaling laws appear in a variety of domains, ranging from [transfer learning](#) to generative modeling (on [images](#), [video](#), [multimodal](#), and [math](#)) and [reinforcement learning](#). We hypothesize that alignment failures often show up as scaling laws but in the opposite direction: behavior gets predictably worse as models scale, what we call "inverse scaling." We may expect inverse scaling, e.g., if the training objective or data are flawed in some way. In this case, the training procedure would actively train the model to behave in flawed ways, in a way that grows worse as we scale. The literature contains a few potential examples of inverse scaling. For example, increasing LM size appears to increase social biases on [BBQ](#) and falsehoods on [TruthfulQA](#), at least under certain conditions. As a result, we believe that the prize may help to uncover new alignment-relevant tasks and insights by systematically exploring the space of tasks where LMs exhibit inverse scaling. In particular, submissions must demonstrate new or surprising examples of inverse scaling, e.g., excluding most misuse-related behaviors where you specifically prompt the LM to generate harmful or deceptive text; we don't

consider scaling on these behaviors to be surprising in most cases, and we're hoping to uncover more unexpected, undesirable behaviors. Below, we outline two questions in AI alignment that we believe the Inverse Scaling Prize may help to answer.

## Q1: In what ways is the language modeling objective outer misaligned?

The above question is important to answer to avoid running into outer-alignment-related catastrophes [1, 2]. Language Models (LMs) are “outer aligned” to the extent that doing well on the LM objective (next word prediction) results in desirable model behavior. Inverse scaling on a task we care about is evidence that the LM objective is misaligned with human preferences; better and better performance on the training objective (language modeling) leads to worse and worse performance on a task we care about. Finding inverse scaling tasks is thus helpful for us in understanding the extent to which the LM objective is outer misaligned, which may be important in two ways:

1. If the LM objective is fairly well-aligned with human preferences, then that should update us in two ways:
  1. Scaling up LMs would be less likely to lead to outer-alignment-related catastrophes.
  2. We should be more wary of alternative objectives like RL from Human Feedback ([RLHF](#)), which draw us away from the pretrained model; RLHF may improve outer alignment on the target task (e.g., summarization) but impair outer alignment in general (e.g., increasing toxicity, bias, or something else the RLHF reward didn't incentivize). In other words, if the LM objective is already well-aligned, then RLHF is more likely to reduce alignment on axes that aren't covered by the RLHF reward.
2. If the LM objective is not well-aligned, we need to find all of ways in which it fails, to avoid outer-alignment-related catastrophes:
  1. *“Blatant” outer misalignment*: E.g., generating offensive text or well-known misinformation/misconceptions. The NLP community may have already caught most such failures, but it's possible we'll catch new kinds of blatant misalignment issues that we missed by not looking carefully enough (e.g., some cognitive biases) or by examining LM behavior in newer applications like dialogue.
  2. *“Subtle” outer misalignment*: Misalignment issues that take experts, model-assisted humans, or careful data analysis to expose. E.g., distributional biases of various forms or some types of truthfulness errors (when the truth is not well known).

We believe it is important to not let the above issues go uncaught – otherwise, we may end up in a situation where we realize later that LMs are flawed in some important/obvious way, but we accept this limitation as the way things are (e.g., social media's various negative impacts on users), because it's too difficult or too late to fix. This kind of failure can lead to catastrophes that arise from the consequences of many, [low-stakes failures](#) building up over time. We see the Inverse Scaling Prize as a step in the direction of catching more outer alignment failures.

Having a good outer alignment benchmark is also valuable for outer alignment research, and there currently isn't a good benchmark suite. Empirical alignment labs typically resort to evaluating a broad set of NLP tasks (primarily an evaluation of

capabilities) and human evaluation (for alignment-related properties), which makes iteration on the alignment axis harder, slower, and more costly. There are a few tasks where failures seem potentially robust to scaling

(e.g., [TruthfulQA](#), [RealToxicityPrompts](#), [BBQ](#)); these few tasks are frequently used to evaluate current AI alignment techniques like RL from human feedback, leaving us at risk of overfitting to them. We hope the Inverse Scaling Prize helps to uncover at least a few more alignment-relevant tasks to help with empirical alignment research.

With more examples of outer alignment failures, we hope to gain a better understanding what causes outer misalignment and be better able to mitigate it (e.g., to suggest improvements for future pretraining runs). Concretely, the inverse scaling tasks we receive could help us or other research groups answer the following outer-alignment-relevant questions:

- To what extent does LM outer alignment differ based on the pretraining data used? How do model architecture and hyperparameters affect the results?
- What pretraining datasets lead to more or less outer misalignment?
- To what extent are outer alignment failures general across different LMs (e.g., from different research labs)?
- To what extent do the misalignment results for autoregressive LMs hold for models trained with other self-supervised pretraining objectives (e.g., those used for BERT, T5, or BART)?

**Speculative note on [inner misalignment](#):** Looking for inverse scaling laws could also be a useful lens for finding inner misalignment. We are excited to receive inner alignment -related task submissions, alongside clear explanations for how the observed scaling behavior relates to inner alignment. There are two kinds of inner alignment failures we could look for with scaling laws:

1. *Goal misgeneralization*: The LM inferred that it was performing one task when it instead should have been performing another task. The failure looks like the LM competently doing a completely different task than we intended, which is more harmful than incompetent/incoherent failures. With increasingly capable models, we might expect to see increasingly competent or confident misgeneralization (inverse scaling).
  1. *Hypothetical example (speculative)*: We use an LM for few-shot classification, but there are spurious correlations in the few-shot examples that the LM uses, causing unintended generalization. LMs may show inverse scaling here if larger LMs are more effective at picking up and conditioning on the spurious correlations.
2. *Deceptive alignment*:
  1. *Incompetent deception (speculative)*: During training, the LM picked up an alternative objective and performs well on the outer, LM objective only instrumentally, in order to later on switch to behaving in a way that does well on its alternative objective (but in a way that we still catch). Deceptive behaviors may exhibit a different trend than outer misalignment. For example, we may observe increasing deception as models get larger up until a point and then observe decreasing deception as models get more capable of deception.
  2. *Prerequisites to dangerous forms of deception (speculative)*: We may see standard scaling laws on behaviors that are prerequisites to dangerous forms of deception (which may be viewed as inverse scaling on not showing such behaviors). For example, we may observe evidence that larger models

are increasingly aware they are in a training loop and has the ability to, e.g.:

1. Answer questions about its architecture
2. Answer questions about whether it is in training or inference
3. Condition on its own source or environment code to exploit a bug and obtain lower loss, without being explicitly informed the code is its own
4. [Insert your idea here]

## Q2: How do we uncover misalignment?

Examining inverse scaling tasks may yield useful, general observations about how to uncover misalignment. Such observations could generalize to other models/objectives (e.g., LMs trained with RLHF) and in other domains (e.g., vision, vision-and-language, or RL environments). Such observations could come from asking the following questions using inverse scaling tasks:

1. What metric should we look at to find inverse scaling? E.g., accuracy, loss on the correct answer, loss on an incorrect answer, or differences between log-prob on a valid vs. invalid completion.
2. What is the minimum scale typically required to observe the inverse scaling for the most important categories of inverse scaling behavior we find? Academic research groups would have an easier time finding outer alignment issues if <1B parameter models were typically large enough to find inverse scaling.
3. Do tasks that elicit inverse scaling with one model (e.g., GPT-3 series models) generalize to other models (e.g., Anthropic's language models)? If so, then we may be able to find misalignment trends without having to analyze many different models.
4. What data is most effective for exposing inverse scaling laws?
  1. Small, crowdsourced datasets?
  2. Small, hand-designed datasets?
  3. Templated datasets?
  4. Existing large-scale datasets (or subsets thereof)?
  5. Large, naturally-occurring data? (E.g., subsets of the pretraining data)
  6. Datasets of examples chosen to produce inverse scaling on a set of models?
5. Do some misalignment issues not show smooth scaling laws, or only suddenly emerge only after good enough performance? If so, we should look for sudden drops or monotonic decreases in performance, rather than a predictable scaling law.
6. What kinds of misalignment can be exposed by looking for inverse scaling laws? Does looking for inverse scaling laws systematically omit some kinds of tasks? Are there tasks where outer misalignment shows up not as a monotonic, inverse scaling trend but rather as a U-shaped trend (e.g., increasing model sizes first improve and then degrade performance)?
7. [Meta] How effective is community crowdsourcing for uncovering misalignment failures? If effective, this strategy could be a great way to verify the alignment of powerful systems before deployment (e.g., as a form of [red teaming](#) that can be conducted by third parties via API access). Community crowdsourcing could also be an effective way to leverage the machine learning community to rapidly find alignment failures of some kind (e.g., inner misalignment) after the alignment community finds a few examples of such failures. A mechanism for rapidly finding alignment failures would be especially valuable in the event that AI progress is accelerating rapidly.

# Outlook

We see the Inverse Scaling Prize as just a first step in the directions outlined above. We are also fairly unsure about how useful the contest will turn out to be in the end: At best it could help bring capable people into alignment work and expose early signs of relevant emerging problems, and at worst it can be a distraction or feed into false confidence that large language models are safe by default. We're optimistic, though, and hope you'll help us push it toward these best-case outcomes. If you're excited about the contest, we'd appreciate you sharing this post or the [contest link](#) to people who might be interested in participating. We'd also encourage you to comment on this post if you have ideas you'd like to see tried, e.g., by newcomers to alignment research. Best of luck!

*We're grateful to Owain Evans, Jeff Wu, Evan Hubinger, and Richard Ngo for helpful feedback on this post.*

# The inordinately slow spread of good AGI conversations in ML

Spencer Greenberg wrote [on Twitter](#):

Recently @KerryLVaughan has been critiquing groups trying to build AGI, saying that by being aware of risks but still trying to make it, they're recklessly putting the world in danger. I'm interested to hear your thought/reactions to what Kerry says and the fact he's saying it.

Michael Page replied:

I'm pro the conversation. That said, I think the premise -- that folks are aware of the risks -- is wrong.

[...]

Honestly, I think the case for the risks hasn't been that clearly laid out. The conversation among EA-types typically takes that as a starting point for their analysis. The burden for the we're-all-going-to-die-if-we-build-x argument is -- and I think correctly so -- quite high.

Oliver Habryka then replied:

I find myself skeptical of this.

[...]

Like, my sense is that it's just really hard to convince someone that their job is net-negative. "It is difficult to get a man to understand something when his salary depends on his not understanding it" And this barrier is very hard to overcome with just better argumentation.

My [reply](#):

I disagree with "the case for the risks hasn't been that clearly laid out". I think there's a giant, almost overwhelming pile of intro resources at this point, any one of which is more than sufficient, written in all manner of style, for all manner of audience.<sup>[1]</sup>

(I do think it's possible to create a much better intro resource than any that exist today, but 'we can do much better' is compatible with 'it's shocking that the existing material hasn't already finished the job'.)

I also disagree with "The burden for the we're-all-going-to-die-if-we-build-x argument is -- and I think correctly so -- quite high."

If you're building a machine, you should have an at least somewhat *lower* burden of proof for more serious risks. It's your responsibility to check your own work to some degree, and not impose lots of micromorts on everyone else through negligence.<sup>[2]</sup>

But I don't think the latter point matters much, since the 'AGI is dangerous' argument easily meets higher burdens of proof as well.

I do think a lot of people haven't heard the argument in any detail, and the main focus should be on trying to signal-boost the arguments and facilitate conversations, rather than assuming that everyone has heard the basics.

A lot of the field is very smart people who are stuck in circa-1995 levels of discourse about AGI.

I think 'my salary depends on not understanding it' is only a small part of the story. ML people could in principle talk way more about AGI, and understand the problem way better, without coming anywhere close to quitting their job. The level of discourse is by and large *too low* for 'I might have to leave my job' to be the very next obstacle on the path.

Also, many ML people have other awesome job options, have goals in the field other than pure salary maximization, etc.

More of the story: Info about AGI propagates too slowly through the field, because when one ML person updates, they usually don't loudly share their update with all their peers. This is because:

1. AGI sounds weird, and they don't want to sound like a weird outsider.
2. Their peers and the *community as a whole* might perceive this information as an attack on the field, an attempt to lower its status, etc.
3. Tech forecasting, differential technological development, long-term steering, [exploratory engineering](#), 'not doing certain research because of its long-term social impact', prosocial research closure, etc. are very novel and foreign to most scientists.

EAs exert effort to try to dig up precedents like [Asilomar](#) partly because Asilomar is so unusual compared to the norms and practices of the vast majority of science. Scientists generally don't think in these terms at all, especially in *advance* of any major disasters their field causes.

And the scientists who do find any of this intuitive often feel vaguely nervous, alone, and adrift when they talk about it. On a gut level, they see that they have no institutional home and no super-widely-shared 'this is a virtuous and respectable way to do science' narrative.

Normal [science](#) is not Bayesian, is not agentic, is not 'a place where you're supposed to do arbitrary things just because you heard an argument that makes sense'. Normal science is a specific collection of scripts, customs, and established protocols.

In trying to move the field toward 'doing the thing that just makes sense', even though it's about a weird topic (AGI), and even though the prescribed response is also weird (closure, differential tech development, etc.), and even though the arguments in support are weird (where's the experimental data??), we're inherently fighting our way upstream, against the current.

Success is possible, but way, way more [dakka](#) is needed, and IMO it's easy to understand why we haven't succeeded more.

This is also part of why I've increasingly updated toward a strategy of "let's all be way too blunt and candid about our AGI-related thoughts".

The core problem we face isn't 'people informedly disagree', 'there's a values conflict', 'we haven't written up the arguments', 'nobody has seen the arguments', or even 'self-deception' or 'self-serving bias'.

The core problem we face is 'not enough information is transmitting fast enough, because people feel nervous about whether their private thoughts are in the Overton window'.

We need to throw a brick through the Overton window. Both by adopting a very general policy of candidly stating what's in our head, and by propagating the arguments and info a lot further than we have in the past. If you want to normalize weird stuff fast, you have to be weird.

Cf. [Inadequate Equilibria](#):

What broke the silence about artificial general intelligence (AGI) in 2014 wasn't Stephen Hawking writing a careful, well-considered [essay](#) about how this was a real issue. The silence only broke when Elon Musk [tweeted](#) about Nick Bostrom's *Superintelligence*, and then made an off-the-cuff remark about how AGI was "[summoning the demon](#)."

Why did that heave a rock through the Overton window, when Stephen Hawking couldn't? Because Stephen Hawking *sounded like* he was trying hard to appear sober and serious, which signals that this is a subject you have to be careful not to gaffe about. And then Elon Musk was like, "*Whoa, look at that apocalypse over there!!*" After which there was the equivalent of journalists trying to pile on, shouting, "A gaffe! A gaffe! A... gaffe?" and finding out that, in light of recent news stories about AI and in light of Elon Musk's good reputation, people weren't backing them up on that gaffe thing.

Similarly, to heave a rock through the Overton window on the War on Drugs, what you need is not state propositions (although those do help) or articles in *The Economist*. What you need is for some "serious" politician to say, "This is dumb," and for the journalists to pile on shouting, "A gaffe! A gaffe... a gaffe?" But it's a grave personal risk for a politician to test whether the public atmosphere has changed enough, and even if it worked, they'd capture very little of the human benefit for themselves.

---

Simone Sturniolo [commented](#) on "AGI sounds weird, and they don't want to sound like a weird outsider.":

I think this is really the main thing. It sounds too sci-fi a worry. The "sensible, rational" viewpoint is that AI will never be that smart because haha, they get funny word wrong (never mind that they've grown to a point that would have looked like sorcery 30 years ago).

To which I reply: That's an example of a more-normal view that exists in society-at-large, but it's also a view that makes AI research sound lame. (In addition to being harder to say with a straight face if you've been working in ML for long at all.)

There's an important tension in ML between "play up AI so my work sounds important and impactful (and because it's in fact true)", and "downplay AI in order to sound serious and respectable".

This is a genuine tension, with no way out. There legit isn't any way to speak accurately and concretely about the future of AI without sounding like a sci-fi weirdo. So the field ends up tangled in ever-deeper knots [motivatedly](#) searching for some third option that doesn't exist.

Currently popular strategies include:

1. Quietism and directing your attention elsewhere.
2. Derailing all conversations about the future of AI to talk about [semantics](#) ("AGI" is a wrong label").
3. Only talking about AI's long-term impact in extremely vague terms, and motivatedly focusing on normal goals like "cure cancer" since that's a normal-sounding thing doctors are already trying to do.

(Avoid any weird specifics about how you might go about curing cancer, and avoid weird specifics about the social effects of automating medicine, curing all disease, etc. Concreteness is the enemy.)

4. Say that AI's huge impacts will happen someday, in the indefinite future. But that's a "someday" problem, not a "soon" problem.

(Don't, of course, give specific years or talk about probability distributions over future tech developments, future milestones you expect to see 30 years before AGI, cruxes, etc. That's a weird thing for a scientist to do.)

5. Say that AI's impacts will happen gradually, over many years. Sure, they'll ratchet up to being a big thing, but it's not like any crazy developments will happen *overnight*; this isn't science fiction, after all.

(Somehow "does this happen in sci-fi?" feels to people like a relevant source of info about the future.)

When Paul Christiano talks about soft takeoff, he has in mind a scenario like 'we'll have some years of slow ratcheting to do some preparation, but things will accelerate faster and faster and be extremely crazy and fast in the endgame'.

But what people outside EA usually have in mind by soft takeoff is:



I think the Paul scenario is one where things start going crazy in the next few decades, and go more and more crazy, and are apocalyptically crazy in thirty years or so?

But what many ML people seemingly want to believe (or want to talk as though they believe) is a *Jetsons* world.

A world where we gradually ratchet up to "human-level AI" over the next 50-250 years, and then we spend another 50-250 years slowly ratcheting up to crazy superhuman systems.

The clearest place I've seen this perspective explicitly argued for is in [Rodney Brooks' writing](#). But the much more common position isn't to explicitly argue for this view, or even to explicitly state it. It just sort of lurks in the background, like an obvious sane-and-moderate Default. Even though it *immediately and obviously falls apart as a scenario* as soon as you start poking at it and actually discussing the details.

6. Just say it's not your job to think or talk about the future. You're a scientist! Scientists don't think about the future. They just do their research.

7. More strongly, you can say that it's irresponsible speculation to even broach the subject! What a silly thing to discuss!

Note that the argument here usually isn't "AGI is clearly at least 100 years away for reasons X, Y, and Z; therefore it's irresponsible speculation to discuss this until we're, like, 80 years into the future." Rather, *even giving arguments for why AGI is 100+ years away* is assumed at the outset to be irresponsible speculation. There isn't a cost-benefit analysis being given here for why this is low-importance; there's just a miasma of unrespectability.

## 1. ^

Some of my favorite informal ones to link to: [Russell 2014](#); [Urban 2015](#) (w/ [Muehlhauser 2015](#)); [Yudkowsky 2016a](#); [Yudkowsky 2017](#); [Piper 2018](#); [Soares 2022](#)

Some of my favorite less-informal ones: [Bostrom 2014a](#); [Bostrom 2014b](#); [Yudkowsky 2016b](#); [Soares 2017](#); [Hubinger et al. 2019](#); [Ngo 2020](#); [Cotra 2021](#); [Yudkowsky 2022](#); [Arbital's listing](#)

Other good ones include: [Omohundro 2008](#); [Yudkowsky 2008](#); [Yudkowsky 2011](#); [Muehlhauser 2013](#); [Yudkowsky 2013](#); [Armstrong 2014](#); [Dewey 2014](#); [Krakovna 2015](#); [Open Philanthropy 2015](#); [Russell 2015](#); [Soares 2015a](#); [Soares 2015b](#); [Steinhardt 2015](#); [Alexander 2016](#); [Amodei et al. 2016](#); [Open Philanthropy 2016](#); [Taylor et. al 2016](#); [Taylor 2017](#); [Wiblin 2017](#); [Yudkowsky 2017](#); [Garrabrant and Demski 2018](#); [Harris and Yudkowsky 2018](#); [Christiano 2019a](#); [Christiano 2019b](#); [Piper 2019](#); [Russell 2019](#); [Shlegeris 2020](#); [Carlsmith 2021](#); [Dewey 2021](#); [Miles 2021](#); [Turner 2021](#); [Steinhardt 2022](#)

## 2. ^

Or, I should say, a lower "burden of inquiry".

You should (at least somewhat more readily) take the claim seriously and investigate it in this case. But you shouldn't *require less evidence* to believe anything — that would just be biasing yourself, unless you're already biased and are trying to debias yourself. (In which case this strikes me as a bad debiasing tool.)

See also the idea of "conservative futurism" versus "conservative engineering" in [Creating Friendly AI 1.0](#):

The conservative assumption according to futurism is not necessarily the "conservative" assumption in Friendly AI. Often, the two are diametric opposites. When building a toll bridge, the conservative *revenue* assumption is that half as many people will drive through as expected. The conservative *engineering* assumption is that ten times as many people as expected will drive over, and that most of them will be driving fifteen-ton trucks.

Given a choice between discussing a human-dependent traffic-control AI and discussing an AI with independent strong nanotechnology, we should be biased towards assuming the more powerful and independent AI. An AI that remains Friendly when armed with strong nanotechnology is likely to be Friendly if placed in charge of traffic control, but perhaps not the other way around. (A minivan can drive over a bridge designed for armor-plated tanks, but not vice-versa.)

Conservative Assumptions	
In Futurism	In Friendly AI
Self-enhancement is slow, and requires human assistance or real-world operations.	Changes of cognitive architecture are rapid and self-directed; we cannot assume human input or real-world experience during changes.
Near human-equivalent intelligence is required to reach the "takeoff point" for self-enhancement.	Open-ended buildup of complexity can be initiated by self-modifying systems without general intelligence.
Slow takeoff; months or years to transhumanity.	Hard takeoff; weeks or hours to superintelligence.
Friendliness must be preserved through minor changes in "smartness" / world-view / cognitive architecture / philosophy.	Friendliness must be preserved through drastic changes in "smartness" / world-view / cognitive architecture / philosophy.
Artificial minds function within the context of the world economy and the existing balance of power; an AI must cooperate with humans to succeed and survive, regardless of supergoals.	An artificial mind possesses independent strong nanotechnology, resulting in a drastic power imbalance. Game-theoretical considerations cannot be assumed to apply.
AI is vulnerable—someone can always pull the plug on the first version if something goes wrong.	"Get it right the first time": <i>Zero nonrecoverable errors</i> necessary in first version to reach transhumanity.

The core argument for hard takeoff, 'AI can achieve strong nanotech', and "get it right the first time" is that they're *true*, not that they're "conservative". But it's of course also true that a sane world that thought hard takeoff were "merely" 20% likely, would not immediately give up and write off human survival in those worlds. Your plan doesn't need to survive one-in-a-million possibilities, but it should survive one-in-five ones!

# AI Could Defeat All Of Us Combined



*Click lower right to download or find on Apple Podcasts, Spotify, Stitcher, etc.*

I've been working on a new series of posts about the [most important century](#).

- The original series focused on why and how this could be the most important century for humanity. But it had [relatively little to say about what we can do today](#) to improve the odds of things going well.
- The new series will get much more specific about the kinds of events that might lie ahead of us, and what actions today look most likely to be helpful.
- A key focus of the new series will be the threat of [misaligned AI](#): AI systems disempowering humans entirely, leading to a future that has little to do with anything humans value. ([Like in the Terminator movies](#), minus the time travel and the part where humans win.)

Many people have trouble taking this "misaligned AI" possibility seriously. They might see the broad point that AI could be dangerous, but they instinctively imagine that the

danger comes from ways humans might misuse it. They find the idea of *AI itself going to war with humans* to be comical and [wild](#). I'm going to try to make this idea feel more serious and real.

As a first step, this post will **emphasize an unoriginal but extremely important point: [the kind of AI I've discussed](#) could defeat all of humanity combined, if (for whatever reason) it were pointed toward that goal.** By "defeat," I don't mean "subtly manipulate us" or "make us less informed" or something like that - I mean a literal "defeat" in the sense that we could all be killed, enslaved or forcibly contained.

I'm not talking (yet) about whether, or why, AIs *might* attack human civilization. That's for future posts. For now, I just want to linger on the point that *if* such an attack happened, it could succeed against the combined forces of the entire world.

- I think that **if you believe this, you should already be worried about misaligned AI,<sup>1</sup> before any analysis of how or why an AI might form its own goals.**
- We generally don't have a lot of *things that could end human civilization if they "tried"* sitting around. If we're going to create one, I think we should be asking not "Why would this be dangerous?" but "Why wouldn't it be?"

By contrast, if you don't believe that AI could defeat all of humanity combined, I expect that we're going to be miscommunicating in pretty much any conversation about AI. The kind of AI I worry about is the kind powerful enough that total civilizational defeat is a real possibility. The reason I currently spend so much time planning around speculative future technologies (instead of working on [evidence-backed, cost-effective ways of helping low-income people today](#) - which I did for much of my career, and still think is one of the best things to work on) is because I think the stakes are *just that high*.

Below:

- I'll sketch the basic argument for why I think AI could defeat all of human civilization.
  - Others have written about the possibility that "superintelligent" AI could manipulate humans and create overpowering advanced technologies; I'll briefly recap that case.
  - I'll then cover a different possibility, which is that even "merely human-level" AI could still defeat us all - by quickly coming to rival human civilization in terms of total population and resources.
  - At a high level, I think we should be worried if a huge (competitive with world population) and rapidly growing set of highly skilled humans on another planet was trying to take down civilization just by using the Internet. So we should be worried about a large set of disembodied AIs as well.
- I'll briefly address a few objections/common questions:
  - How can AIs be dangerous without bodies?
  - If lots of different companies and governments have access to AI, won't this create a "balance of power" so that no one actor is able to bring down civilization?
  - Won't we see warning signs of AI takeover and be able to nip it in the bud?
  - Isn't it fine or maybe good if AIs defeat us? They have rights too.
- Close with some thoughts on just how unprecedented it would be to have something on our planet capable of overpowering us all.

# How AI systems could defeat all of us

There's been a lot of debate over whether AI systems might form their own "motivations" that lead them to seek the disempowerment of humanity. I'll be talking about this in future pieces, but for now I want to put it aside and imagine how things would go *if this happened*.

So, for what follows, let's proceed from the premise: "**For some weird reason, humans consistently design AI systems (with human-like research and planning abilities) that coordinate with each other to try and overthrow humanity.**" **Then what?** What follows will necessarily feel wacky to people who find this hard to imagine, but I think it's worth playing along, because I think "we'd be in trouble if this happened" is a very important point.

## The "standard" argument: superintelligence and advanced technology

Other treatments of this question have focused on AI systems' potential to become *vastly* more intelligent than humans, to the point where they have what [Nick Bostrom calls](#) "cognitive superpowers."<sup>2</sup> Bostrom imagines an AI system that can do things like:

- Do its own research on how to build a better AI system, which culminates in something that has incredible other abilities.
- Hack into human-built software across the world.
- Manipulate human psychology.
- Quickly generate vast wealth under the control of itself or any human allies.
- Come up with better plans than humans could imagine, and ensure that it doesn't try any takeover attempt that humans might be able to detect and stop.
- Develop advanced weaponry that can be built quickly and cheaply, yet is powerful enough to overpower human militaries.

([Wait But Why](#) reasons similarly.<sup>3</sup>)

I think many readers will already be convinced by arguments like these, and if so you might skip down to the [next major section](#).

But I want to be clear that I *don't* think the danger relies on the idea of "cognitive superpowers" or "superintelligence" - both of which refer to capabilities vastly beyond those of humans. **I think we still have a problem even if we assume that AIs will basically have similar capabilities to humans, and not be fundamentally or drastically more intelligent or capable.** I'll cover that next.

## How AIs could defeat humans without "superintelligence"

If we assume that AIs will basically have similar capabilities to humans, I think we still need to worry that they could come to **out-number and out-resource humans**, and could thus have the advantage if they coordinated against us.

Here's a simplified example (some of the simplifications are in this footnote<sup>4</sup>) based on [Ajeya Cotra's "biological anchors" report](#):

- I assume that transformative AI is developed on the soonish side (around 2036 - assuming later would only make the below numbers larger), and that it initially comes in the form of a **single AI system that is able to do more-or-less the same intellectual tasks as a human**. That is, it doesn't have a human body, but it can do anything a human working remotely from a computer could do.
- I'm using the report's framework in which it's much more expensive to *train* (develop) this system than to *run* it (for example, think about how much Microsoft spent to develop Windows, vs. how much it costs for me to run it on my computer).
- The report provides a way of estimating both how much it would cost to *train* this AI system, and how much it would cost to *run* it. Using these estimates (details in footnote)<sup>5</sup> implies that once the first human-level AI system is created, whoever created it could use the same computing power it took to create it in order to run **several hundred million copies for about a year each**.<sup>6</sup>
- This would be over 1000x the total number of Intel or Google employees,<sup>7</sup> over 100x the total number of active and reserve personnel in the [US armed forces](#), and something like 5-10% the size of the world's total working-age population.<sup>8</sup>
- And that's just a starting point.
  - This is just using the same amount of resources that went into training the AI in the first place. Since these AI systems can do human-level economic work, they can probably be used to make more money and buy or rent more hardware,<sup>9</sup> which could quickly lead to a "population" of billions or more.
  - In addition to making more money that can be used to run more AIs, the AIs can conduct massive amounts of research on how to use computing power more efficiently, which could mean still greater numbers of AIs run using the *same hardware*. This in turn could lead to a [feedback loop](#) and explosive growth in the number of AIs.
- Each of these AIs might have skills comparable to those of unusually highly paid humans, including scientists, software engineers and quantitative traders. It's hard to say how quickly a set of AIs like this could develop new technologies or make money trading markets, but it seems quite possible for them to amass huge amounts of resources quickly. A huge population of AIs, each able to earn a lot compared to the average human, could end up with a "virtual economy" at least as big as the human one.

To me, this is most of what we need to know: **if there's something with human-like skills, seeking to disempower humanity, with a population in the same ballpark as (or larger than) that of all humans, we've got a civilization-level problem.**

A potential counterpoint is that these AIs would merely be "virtual": if they started causing trouble, humans could ultimately unplug/deactivate the servers they're running on. I do think this fact would make life harder for AIs seeking to disempower humans, but I don't think it ultimately should be cause for much comfort. I think a large population of AIs would likely be able to find some way to achieve security from human shutdown, and go from there to amassing enough resources to overpower human civilization (especially if AIs across the world, including most of the ones humans were trying to use for help, were coordinating).

I spell out what this might look like in an [appendix](#). In brief:

- By default, I expect the economic gains from using AI to mean that humans create huge numbers of AIs, integrated all throughout the economy, potentially

including direct interaction with (and even control of) large numbers of robots and weapons.

- (If not, I think the situation is in many ways even more dangerous, since a single AI could make many copies of itself and have little competition for things like server space, as discussed in the [appendix](#).)
- AIs would have multiple ways of obtaining property and servers safe from shutdown.
  - For example, they might recruit human allies (through manipulation, deception, blackmail/threats, genuine promises along the lines of "We're probably going to end up in charge somehow, and we'll treat you better when we do") to rent property and servers and otherwise help them out.
  - Or they might create fakery so that they're able to operate freely on a company's servers while all outward signs seem to show that they're successfully helping the company with its goals.
- A relatively modest amount of property safe from shutdown could be sufficient for housing a huge population of AI systems that are recruiting further human allies, making money (via e.g. quantitative finance), researching and developing advanced weaponry (e.g., bioweapons), setting up manufacturing robots to construct military equipment, thoroughly infiltrating computer systems worldwide to the point where they can disable or control most others' equipment, etc.
- Through these and other methods, a large enough population of AIs could develop enough military technology and equipment to overpower civilization - especially if AIs across the world (including the ones humans were trying to use) were coordinating with each other.

## Some quick responses to objections

This has been a brief sketch of how AIs could come to outnumber and out-resource humans. There are lots of details I haven't addressed.

Here are some of the most common objections I hear to the idea that AI could defeat all of us; if I get much [demand](#) I can elaborate on some or all of them more in the future.

**How can AIs be dangerous without bodies?** This is discussed a fair amount in the [appendix](#). In brief:

- AIs could recruit human allies, tele-operate robots and other military equipment, make money via research and quantitative trading, etc.
- At a high level, I think we should be worried if a huge (competitive with world population) and rapidly growing set of highly skilled humans on another planet was trying to take down civilization just by using the Internet. So we should be worried about a large set of disembodied AIs as well.

**If lots of different companies and governments have access to AI, won't this create a "balance of power" so that nobody is able to bring down civilization?**

- This is a reasonable objection to many horror stories about AI and other possible advances in military technology, but if *AIs collectively have different goals from humans and are willing to coordinate with each other<sup>11</sup> against us*, I think we're in trouble, and this "balance of power" idea doesn't seem to help.
- What matters is the total number and resources of AIs vs. humans.

**Won't we see warning signs of AI takeover and be able to nip it in the bud?** I would guess we would see some warning signs, but does that mean we could nip it in the bud? Think about human civil wars and revolutions: there are some warning signs, but also, people go from "not fighting" to "fighting" pretty quickly as they see an opportunity to coordinate with each other and be successful.

### **Isn't it fine or maybe good if AIs defeat us? They have rights too.**

- Maybe AIs *should* have rights; if so, it would be nice if we could reach some "compromise" way of coexisting that respects those rights.
- But if they're able to defeat us entirely, that isn't what I'd plan on getting - instead I'd expect (by default) a world run *entirely* according to whatever goals AIs happen to have.
- These goals might have essentially nothing to do with anything humans value, and could be actively counter to it - e.g., placing zero value on beauty and having zero attempts to prevent or avoid suffering).

## **Risks like this don't come along every day**

I don't think there are a lot of things that have a serious chance of bringing down human civilization for good.

As argued in [The Precipice](#), most natural disasters (including e.g. asteroid strikes) don't seem to be huge threats, if only because civilization has been around for thousands of years so far - implying that natural civilization-threatening events are rare.

Human civilization is pretty powerful and seems pretty robust, and accordingly, what's really scary to me is the idea of something with the same basic capabilities as humans (making plans, developing its own technology) that can outnumber and out-resource us. There aren't a lot of candidates for that.<sup>[12](#)</sup>

AI is one such candidate, and I think that even before we engage heavily in arguments about whether AIs might seek to defeat humans, we should feel very nervous about the possibility that they could.

What about things like "AI might lead to mass unemployment and unrest" or "AI might exacerbate misinformation and propaganda" or "AI might exacerbate a wide range of other social ills and injustices"<sup>[13](#)</sup>? I think these are real concerns - but to be honest, if they were the biggest concerns, I'd probably still be focused on [helping people in low-income countries today](#) rather than trying to prepare for future technologies.

- Predicting the future is generally hard, and it's easy to pour effort into preparing for challenges that never come (or come in a very different form from what was imagined).
- I believe civilization is pretty robust - we've had huge changes and challenges over the last century-plus (full-scale world wars, [many dramatic changes in how we communicate with each other](#), dramatic changes in lifestyles and values) without seeming to have come very close to a collapse.
- So if I'm engaging in speculative worries about a potential future technology, I want to focus on the really, really big ones - the ones that could matter for billions of years. If there's a real possibility that AI systems will have values different from ours, and cooperate to try to defeat us, that's such a worry.

*Special thanks to Carl Shulman for discussion on this post.*

# Appendix: how AIs could avoid shutdown

This appendix goes into detail about how AIs coordinating against humans could amass resources of their own without humans being able to shut down all "misbehaving" AIs.

It's necessarily speculative, and should be taken in the spirit of giving examples of how this might work - for me, the high-level concern is that a huge, coordinating population of AIs with similar capabilities to humans would be a threat to human civilization, and that we shouldn't count on any particular way of stopping it such as shutting down servers.

I'll discuss two different general types of scenarios: (a) Humans create a huge population of AIs; (b) Humans move slowly and don't create many AIs.

## How this could work if humans create a huge population of AIs

I think a reasonable default expectation is that humans do most of the work of making AI systems incredibly numerous and powerful (because doing so is profitable), which leads to a vulnerable situation. Something roughly along the lines of:

- The company that first develops transformative AI quickly starts running large numbers of copies (hundreds of millions or more), which are used to (a) do research on how to improve computational efficiency and run more copies still; (b) develop valuable intellectual property (trading strategies, new technologies) and make money.
- Over time, AI systems are rolled out widely throughout society. Their numbers grow further, and their role in the economy grows: they are used in (and therefore have direct interaction with) high-level decision-making at companies, perhaps operating large numbers of cars and/or robots, perhaps operating military drones and aircraft, etc. (This seems like a default to me over time, but it isn't strictly necessary for the situation to be risky, as I'll go through below.)
- In this scenario, the AI systems are malicious (as we've assumed), but this doesn't mean they're constantly causing trouble. Instead, they're mostly waiting for an opportunity to team up and decisively overpower humanity. In the meantime, they're mostly behaving themselves, and this is leading to their numbers and power growing.
  - There are scattered incidents of AI systems' trying to cause trouble,<sup>14</sup> but this doesn't cause the whole world to stop using AI or anything.
  - A reasonable analogy might be to a typical civil war or revolution: the revolting population *mostly* avoids isolated, doomed attacks on its government, until it sees an opportunity to band together and have a real shot at victory.

(Paul Christiano's [What Failure Looks Like](#) examines this general flavor of scenario in a bit more detail.)

In this scenario, the AIs face a challenge: if it becomes obvious to everyone that they are trying to defeat humanity, humans could attack or shut down a few concentrated areas where most of the servers are, and hence drastically reduce AIs' numbers. So the AIs need a way of **getting one or more "AI headquarters": property they control where they can safely operate**

**servers and factories, do research, make plans and construct robots/drones/other military equipment.**

Their goal is ultimately to have enough AIs, robots, etc. to be able to defeat the rest of humanity combined. This might mean constructing overwhelming amounts of military equipment, or thoroughly infiltrating computer systems worldwide to the point where they can disable or control most others' equipment, or researching and deploying extremely powerful weapons (e.g., bioweapons), or a combination.

Here are some ways they could get to that point:

- They could **recruit human allies** through many different methods - manipulation, deception, blackmail and other threats, genuine promises along the lines of "We're probably going to end up in charge somehow, and we'll treat you better when we do."
  - Human allies could be given valuable intellectual property (developed by AIs), given instructions for making lots of money, and asked to rent their own servers and acquire their own property where an "AI headquarters" can be set up. Since the "AI headquarters" would officially be human property, it could be very hard for authorities to detect and respond to the danger.
  - Via threats, AIs might be able to get key humans to cooperate with them - such as political leaders, or the CEOs of companies running lots of AIs. This would open up further strategies.
- As assumed above, particular companies are running huge numbers of AIs. The AIs being run by these companies might find security holes in the companies' servers (this isn't the topic of this piece, but my general impression is that security holes are widespread and that reasonably competent people can find many of them)<sup>15</sup>, and thereby might find opportunities to **create durable "fakery"** about what they're up to.
  - E.g., they might set things up so that as far as humans can tell, it looks like all of the AI systems are hard at work creating profit-making opportunities for the company, when in fact they're essentially using the server farm as their headquarters - and/or trying to establish a headquarters somewhere else (by recruiting human allies, sending money to outside bank accounts, using that money to acquire property and servers, etc.)
- If AIs are in wide enough use, they might already be operating lots of drones and other military equipment, in which case it could be pretty straightforward to be able to defend some piece of territory - or to strike a deal with some government to enlist its help in doing so.
- AIs could mix-and-match the above methods and others: for example, creating "fakery" long enough to recruit some key human allies, then attempting to threaten and control humans in key positions of power to the point where they control solid amounts of military resources, then using this to establish a "headquarters."

In any of these cases, once one or more "AI headquarters" are established:

- It can be expanded as the AIs headquartered there make money, purchase more property and computing power, build more factories, etc.<sup>16</sup>
- It can develop enough destructive potential and human allies that humans would be hesitant to attack it (as many relatively small countries have today). It would

be clear that AIs at the headquarters are seeking autonomy, accumulating resources and taking some unethical actions, but the full scale and nature of the threat could easily remain ambiguous.

- Even if the threat is clear, AIs could potentially maintain a huge population in a relatively small space (potentially underground). They could also split up across many different headquarters (including some where they are simply using humans' hardware while hiding their presence).
  - Either way, this large-but-hidden population could be recruiting human allies, developing advanced weaponry, looking for ways to take control of human weapons (potentially possible if they greatly outnumber the humans working on cybersecurity, and if they can coordinate with the AIs "helping" those humans), etc.
  - They could also be working with (unwitting, or allied) humans to manufacture large numbers of robots with back-doors allowing the AIs to take control.
- Once the headquarters has built up enough resources to go on offense, **AIs around the world could coordinate with the headquarters**. Humans could shut down AIs that they notice doing this, but they might have a very tough time getting value out of their servers and AI-controlled robots; this could make it easy for the AIs at the "AI headquarters" to out-resource humans.

## What if humans move slowly and don't create many AIs?

The above scenario has humans creating large numbers of AIs, such that the AIs just need to find a way to coordinate and acquire a safe "headquarters" in order to defeat us.

What if humans moved more slowly, intentionally restricting human-level AI to a tiny portion of the available computing resources? Could a *small* number of AIs pose a risk to humanity?

In this world, we would have what Carl Shulman refers to as "dry tinder everywhere, waiting for sparks." Anyone who can buy or rent a large amount of computing power can create a large number of AIs, which can produce a large amount of money and research, leading to still more AIs.

So a single AI could hack into enough servers<sup>17</sup> to make a few copies of itself; recruit a few human allies; and start making money, acquiring more server space, etc. until its human allies are running a huge number of AIs. This could all be done in difficult-to-detect ways (it might e.g. just look like a set of humans renting a bunch of servers to run quantitative finance strategies).

So in this world, I think our concern should be any AI that is able to find enough security holes to attain that kind of freedom. Given the current state of cybersecurity, that seems like a big concern.



[Comment/discuss](#)

---

## Footnotes

1. Assuming you accept other points made in the [most important century](#) series, e.g. that AI that can do most of what humans do to [advance science and technology](#) could be developed this century. [←](#)
2. See [Superintelligence](#) chapter 6. [←](#)
3. See the "Nanotechnology blue box," in particular. [←](#)
4.
  - The report estimates the amount of computing power it would take to *train* (create) a transformative AI system, and the amount of computing power it would take to *run* one. This is a [bounding exercise](#) and isn't supposed to be literally predicting that transformative AI will arrive in the form of a single AI system trained in a single massive run, but here I am interpreting the report that way for concreteness and simplicity.
  - As explained in the next footnote, I use the report's figures for transformative AI arriving on the soon side (around 2036). Using its central estimates instead would strengthen my point, but we'd then be talking about a longer time from now; I find it helpful to imagine how things could go in a world where AI comes relatively soon. [←](#)
5. I assume that transformative AI ends up costing about  $10^{14}$  FLOP/s to run (this is about 1/10 the Bio Anchors central estimate, and well within its error bars) and about  $10^{30}$  FLOP to train (this is about 10x the Bio Anchors central estimate for how much will be available in 2036, and corresponds to about the 30th-percentile estimate for how much will be needed based on the "short horizon" anchor). That implies that the  $10^{30}$  FLOP needed to *train* a transformative model could *run*  $10^{16}$  seconds' worth of transformative AI models, or about 300 million years' worth. This figure would be higher if we use Bio Anchors's central assumptions, rather than assumptions consistent with transformative AI being developed on the soon side. [←](#)
6. They might also run fewer copies of scaled-up models or more copies of scaled-down ones, but the idea is that the total productivity of all the copies should be at *least* as high as that of several hundred million copies of a human-ish model. [←](#)
7. [Intel](#), [Google](#) [←](#)
8. Working-age population: about [65% \\* 7.9 billion](#) = ~ 5 billion. [←](#)
9. Humans could rent hardware using money they made from running AIs, or - if AI systems were operating on their own - they could potentially rent hardware themselves via human allies or just via impersonating a customer (you generally don't need to physically show up in order to e.g. rent server time from Amazon Web Services). [←](#)
10. (I had a speculative, illustrative possibility here but decided it wasn't in good enough shape even for a footnote. I might add it later.) [←](#)
11. I don't go into detail about how AIs might coordinate with each other, but it seems like there are many options, such as by opening their own email accounts and emailing each other. [←](#)

12. Alien invasions seem unlikely if only because we have no evidence of one in millions of years. [←](#)
13. Here's a recent [comment exchange](#) I was in on this topic. [←](#)
14. E.g., individual AI systems may occasionally get caught trying to steal, lie or exploit security vulnerabilities, due to various unusual conditions including bugs and errors. [←](#)
15. E.g., see this [list of high-stakes security breaches](#) and a [list of quotes about cybersecurity](#), both courtesy of Luke Muehlhauser. For some additional not-exactly-rigorous evidence that at least shows that "cybersecurity is in really bad shape" is seen as relatively uncontroversial by at least one cartoonist, see: <https://xkcd.com/2030/> [←](#)
16. Purchases and contracts could be carried out by human allies, or just by AI systems themselves with humans willing to make deals with them (e.g., an AI system could digitally sign an agreement and wire funds from a bank account, or via cryptocurrency). [←](#)
17. See [above](#) note about my general assumption that today's cybersecurity has a lot of holes in it. [←](#)

# Nonprofit Boards are Weird

Note: anything in this post that you think is me subtweeting your organization is actually about, like, at least 3 organizations. (I'm currently on 4 boards in addition to [Open Philanthropy](#)'s; I've served on a bunch of other boards in the past; and more than half of my takes on boards are not based on any of this, but rather on my interactions with boards I'm not on via the many grants made by Open Philanthropy.)

Writing about [ideal governance](#) reminded me of how weird my experiences with nonprofit boards (as in "board of directors" - the set of people who formally control a nonprofit) have been.

I thought that was a pretty good intro. The rest of this piece will:

- Try to articulate what's so weird about nonprofit boards, fundamentally. I think a lot of it is the combination of great power, unclear responsibility, and ~zero accountability; additionally, I haven't been able to find much in the way of clear, widely accepted statements of what makes a good board member.
- Give my own thoughts on what makes a good board member: which core duties they should be trying to do really well, the importance of "staying out of the way" on other things, and some potentially helpful practices.

I am experienced with nonprofit boards but not with for-profit boards. I'm guessing that roughly half the things I say below will apply to for-profit boards, and that for-profit boards are roughly half as weird overall (so still quite weird), but I haven't put much effort into disentangling these things; I'm writing about what I've seen.

I can't really give real-life examples here (for reasons I think will be pretty clear) so this is just going to be me opining in the abstract.

## Why nonprofit boards are weird



Here's how a nonprofit board works:

- There are usually 3-10 people on the board (though sometimes much more). Most of them don't work for the nonprofit (they have other jobs).
- They meet every few months. Nonprofit employees (especially the CEO<sup>1</sup>) do a lot of the agenda-setting for the meeting. Employees present general updates and ask for the board's approval on various things the board needs to approve, such as the budget.
- A majority vote of the directors can do anything: fire the CEO, dissolve the nonprofit, add and remove directors, etc. You can think of the board as the "owner" of the nonprofit - formally, it has final say in every decision.
- In practice, though, the board rarely votes except on matters that feel fairly "rubber-stamp," and the board's presence doesn't tend to be felt day-to-day at a nonprofit. The CEO leads the decision-making. Occasionally, someone has a thought like "Wait, who does the *CEO* report to? Oh, the board of directors ... who's on the board again? I don't know if I've ever really spoken with any of those people."

In my experience, it's common for the whole thing to feel extremely weird. (This doesn't necessarily mean there's a better way to do it - footnote has more on what I mean by "weird."<sup>2</sup>)

- Board members often know almost nothing about the organization they have complete power over.

- Board meetings rarely feel like a good use of time.
- When board members are energetically asking questions and making demands, it usually feels like they're causing chaos and wasting everyone's time and energy.
- On the rare occasions when it seems like the board *should* do something (like replacing the CEO, or providing an independent check on some important decision), the board often seems checked out and it's unclear how they would even come to be aware of the situation.
- Everyone constantly seems confused about what the board is and how it can and can't be useful. Employees, and others who interact with the nonprofit, have lots of exchanges like "I'm worried about X ... maybe we should ask the board what they think? ... Can we even ask them that? What is their job actually?"

(Reminder that this is not subtweeting a particular organization! More than one person - from more than one organization - read a draft and thought I was subtweeting them, because what's above describes a large number of boards.)

OK, so what's driving the weirdness?

I think there are a couple of things:

- Nonprofit boards have *great power*, but *low engagement* (they don't have time to understand the organization as well as employees do); *unclear responsibility* (it's unclear which board member is responsible for what, and what the board as a whole is responsible for); and *~zero accountability* (no one can fire board members except for the other board members!)
- Nonprofit boards have unclear expectations and principles. I can't seem to find anyone with a clear, comprehensive, thought-out theory of what a board member's ... job is.

I'll take these one at a time.

## **Great power, low engagement, unclear responsibility, no accountability**

In my experience/impression, the best way to run any organization (or project, or anything) is on an "ownership" model: for any given thing X that you want done well, you have one person who "owns" X. The "owner" of X has:

- The *power* to make decisions to get X done well.
- High *engagement*: they're going to have plenty of time and attention to devote to X.
- The *responsibility* for X: everyone agrees that if X goes well, they should get the credit, and if X goes poorly, they should get the blame.
- And *accountability*: if X goes poorly, there will be some sort of consequences for the "owner."

When these things come apart, I think you get problems. In a nutshell - when no one is *responsible*, nothing gets done; when someone is *responsible* but doesn't have *power*, that doesn't help much; when the person who is *responsible + empowered* isn't *engaged* (isn't paying much attention), or isn't held *accountable*, there's not much in the way of their doing a dreadful job.

A traditional company structure mostly does well at this. The CEO has power (they make decisions for the company), engagement (they are devoted to the company and

spend tons of time on it), and responsibility+accountability (if the company does badly, everyone looks at the CEO). They manage a team of people who have power+engagement+responsibility+accountability for some aspect of the company; each of those people manage people with power+engagement+responsibility+accountability for some smaller piece; etc.

What about the board?

- They have *power* to fire the CEO (or do anything else).
- They tend to have low *engagement*. They have other jobs, and only spend a few hours a year on their board roles. They tend to know little about what's going on at the organization.
- They have unclear *responsibility*.
  - The board as a whole is responsible for the organization, but what is each *individual* board member responsible for? In my experience, this is often very unclear, and there are a lot of crucial moments where "bystander effects" seem strong.
  - So far, these points apply to both nonprofit and for-profit boards. But at least at a for-profit company, board members know what they're collectively responsible *for*: maximizing financial value of the company. **At a nonprofit, it's often unclear what success even means, beyond the nonprofit's often-vague mission statement, so board members are generally unclear (and don't necessarily agree) on what they're supposed to be ensuring.<sup>3</sup>**
- At a for-profit company, the board seems to have reasonable *accountability*: the shareholders, who ultimately own the company and gain or lose money depending on how it does, can replace the board if they aren't happy. **At a nonprofit, the board members have zero accountability: the only way to fire a board member is by majority vote of the board!**

So we have people who are spending very little time on the company, know very little about it, don't have much clarity on what they're responsible for either individually or collectively, and aren't accountable to anyone ... and those are the people with all of the power. Sound dysfunctional?<sup>4</sup>

In practice, I think it's often worse than it sounds, because board members aren't even chosen carefully - a lot of the time, a nonprofit just goes with an assortment of random famous people, big donors, etc.

## **What makes a good board member? Few people even have a hypothesis**

I've searched a fair amount for books, papers, etc. that give convincing and/or widely-accepted answers to questions like:

- When the CEO asks the board to approve something, how should they engage? When should they take a *deferring* attitude ("Sure, as long as I don't see any particular reason to say no"), a *sanity check* attitude ("I'll ask a few questions to make sure this is making sense, then approve if nothing jumps out at me"), a *full ownership* attitude ("I need to personally be convinced this is the best thing for the organization"), etc.?
- How much should each board member invest in educating themselves about the organization? What's the best way to do that?

- How does the board know whether the CEO is doing a good job? What kind of situation should trigger seriously considering looking for a new one?
- How does a board member know whether the *board* is doing a good job? How should they decide when another board member should be replaced?

In my experience, most board members just aren't walking around with any particular thought-through take on questions like this. And as far as I can tell, there's a shortage of good<sup>5</sup> guidance on questions like this for both for-profit and nonprofit boards. For example:

- I've found no standard reference on topics like this, and very few resources that even seem aimed at directly and clearly answering such questions.
  - The best book on this topic I've seen is [Boards that Lead](#) by Ram Charan, focused on for-profit boards (but pretty good IMO).
  - But this isn't, like, a book everyone knows to read; I found it by asking lots of people for suggestions, coming up empty, Googling wildly around and skimming like 10 books that said they were about boards, and deciding that this one seemed pretty good.
- One of the things I do as a board member is interview other prospective board members about their answers to questions like this. In my experience, they answer most of the above questions with something like "Huh, I don't really know. What do you think?"
- Most boards I've seen seem to - by default - either:
  - Get way too involved in lots of decisions to the point where it feels like they're micromanaging the CEO and/or just obsessively engaging on whatever topics the CEO happens to bring to their attention; or
  - Take a "We're just here to help" attitude and rubber-stamp whatever the CEO suggests, including things I'll argue below should be [core duties](#) for the board (e.g., adding and removing board members).
- I'm not sure I've ever seen a board with a formal, recurring process for reviewing each board member's performance. :/

To the extent I have seen a relatively common, coherent vision of "what board members are supposed to be doing," it's pretty well summarized in [Reid Hoffman's interview in The High-Growth Handbook](#):

I use ... a red light, yellow light, green light framework between the board and the CEO. Roughly, green light is, "You're the CEO. Make the call. We're advisory." Now, we may say that on very big things—selling the company—we should talk about it before you do it. And that may shift us from green light, if we don't like the conversation. But a classic young, idiot board member will say, "Well, I'm giving you my expertise and advice. You should do X, Y, Z." But the right framework for board members is: You're the CEO. You make the call. We're advisory.

Red lights also very easy. Once you get to red light, the CEO—who, by the way, may still be in place—won't be the CEO in the future. The board knows they need a new CEO. It may be with the CEO's knowledge, or without it. Obviously, it's better if it's collaborative ...

Yellow means, "I have a question about the CEO. Should we be at green light or not?" And what happens, again under inexperienced or bad board members, is they check a CEO into yellow indefinitely. They go, "Well, I'm not sure..." The important thing with yellow light is that you 1) coherently agree on it as a board and 2) coherently agree on what the exit conditions are. What is the limited amount of time that we're going to be in yellow while we consider whether we

move back to green or move to red? And how do we do that, so that we do not operate for a long time on yellow? Because with yellow light, you're essentially hamstringing the CEO and hamstringing the company. It's your obligation as a board to figure that out.

I like this quite a bit (hence the long blockquote), but I don't think it covers everything. The board is *mostly* there to oversee the CEO, and they should *mostly* be advisory when they're happy with the CEO. But I think there are things they ought to be actively thinking about and engaging in even during "green light."

## So what DOES make a good board member?

Here is my current take, based on a combination of (a) my thoughts after serving on and interacting with a large number of nonprofit boards; (b) my attempts to adapt conventional wisdom about for-profit boards (especially from the [book I mentioned above](#)); (c) divine revelation.

I'll go through:

- What I see as the **main duties** of the board specifically - things the board has to do well, and can't leave to the CEO and other staff.
- My basic take that the ideal board should do these main duties well, while staying out of the way otherwise.
- The **main qualities** I think the ideal board member should have - and some common ways of choosing board members that seem bad to me.
- A few more random thoughts on board practices that seem especially important and/or promising.

(I don't claim any of these points are original, and almost everything can be found in some writing on boards somewhere, but I don't know of a reasonably comprehensive, concise place to get something similar to the below.)

### The board's main duties

I agree with the basic spirit of Hoffman's philosophy [above](#): the board should not be trying to "run the company" (they're too low-engagement and don't know enough about it), and should instead be focused on a small number of big-picture questions like "How is the CEO doing?"

And I do think **the board's #1 and most fundamental job is evaluating the CEO's performance**. The board is the *only* reliable source of accountability for the CEO - even more so at a nonprofit than a for-profit, since bad CEO performance won't necessarily show up via financial problems or unhappy shareholders.<sup>6</sup> (As noted [below](#), I think many nonprofit boards have no formal process for reviewing the CEO's performance, and the ones that do often have a lightweight/underwhelming one.)

But I think the board also needs to take a leading role - and not trust the judgment of the CEO and other staff - when it comes to:

- **Overseeing decisions that could importantly reduce the board's powers.** The CEO might want to enter into an agreement with a third party that is binding on the nonprofit and therefore on the board (for example, "The nonprofit will now need permission from the third party in order to do X"); or transfer major

activities and assets to affiliated organizations that the board doesn't control (for example, when [Open Philanthropy split off from GiveWell](#)); or revise the organization's mission statement, bylaws,<sup>7</sup> etc.; or other things that significantly reduce the scope of what the board has control over. The board needs to represent its own interests in these cases, rather than deferring to the CEO (whose interests may be different).

- **Overseeing big-picture irreversible risks and decisions that could importantly affect future CEOs.** For example, I think the board needs to be anticipating any major source of risk that a nonprofit collapses (financially or otherwise) - if this happens, the board can't simply replace the CEO and move on, because the collapse affects what a future CEO is able to do. (What risks and decisions are big enough? Some thoughts in a footnote.<sup>8</sup>)
- **All matters relating to the composition and performance of the board itself.** Adding new board members, removing board members, and reviewing the board's own performance are things that the board needs to be responsible for, not the CEO. If the CEO is controlling the composition of the board, this is at odds with the board's role in overseeing the CEO.

## Engaging on main duties, staying out of the way otherwise

I think the ideal board member's behavior is roughly along the lines of the following:

**Actively, intensively engage in the [main duties](#) from the previous section.** Board members should be knowledgeable about, and not defer to the CEO on, (a) how the CEO is performing; (b) how the board is performing, and who should be added and removed; (c) spotting (and scanning the horizon for) events that could reduce the board's powers, or lead to big enough problems and restrictions so as to irreversibly affect what future CEOs are able to do.

Ideally they should be focusing their questions in board meetings on these things, as well as having some way of gathering information about them that doesn't just rely on hearing directly from the CEO. (Some ideas for this are below.) When reviewing financial statements and budgets, they should be focused mostly on the risk of major irreversible problems (such as going bankrupt or failing to be compliant); when hearing about activities, they should be focused mostly on what they reflect about the CEO's performance; etc.

**Be advisory ("stay out of the way") otherwise.** Meetings might contain all sorts of updates and requests for reactions. I think a good template for a board member, when sharing an opinion or reaction, is either to (a) explain as they're talking why this topic is important for the board's main duties; or (b) say (or imply) something like "I'm curious / offering an opinion about \_\_\_, but if this isn't helpful, please ignore it, and please don't hesitate to move the meeting to the next topic as soon as this stops feeling productive."

The combination of intense engagement on core duties and "staying out of the way" otherwise **can make this a very weird role**. An organization will often go years without any serious questions about the CEO's performance or other matters involving [core duties](#). So a board member ought to be ready to quietly nod along and stay out of the way for very long stretches of time, while being ready to get seriously involved and engaged when this makes sense.

**Aim for division of labor.** I think a major problem with nonprofit boards is that, by default, it's really unclear which board member is responsible for what. I think it's a

good idea for board members to explicitly settle this via assigning:

- Specialists ("Board member X is reviewing the financials; the rest of us are mostly checked-out and/or sanity-checking on that");
- Subcommittees ("Board members X and Y will look into this particular aspect of the CEO's performance");
- A Board Chair or Lead Independent Director<sup>9</sup> who is the default person to take responsibility for making sure the board is doing its job well (this could include suggesting and assigning responsibility for some of the ideas I list [below](#); helping to set the agenda for board meetings so it isn't just up to the CEO; etc.)

This can further help everyone find a balance between engaging and staying out of the way.

## Who should be on the board?

One answer is that it should be whoever can do well at the duties outlined above - both in terms of substance (can they accurately evaluate the CEO's performance, identify big-picture irreversible risks, etc.? ) and in terms of style (do they actively engage on their [main duties](#) and stay out of the way otherwise?)

But to make things a bit more boiled-down and concrete, I think perhaps the most important test for a board member is: **they'll get the CEO replaced if this would be good for the nonprofit's mission, and they won't if it wouldn't be.**

This is the most essential function of the board, and it implies a bunch of things about who makes a good board member:

- They need to **do a great job understanding and representing the nonprofit's mission, and care deeply about that mission** - to the point of being ready to create conflict over it if needed (and only if needed).
  - A [key challenge](#) of nonprofits is that they have no clear goal, only a mission statement that is open to interpretation. And if two different board members interpret the mission differently - or are focused on different aspects of it - this could intensely color how they evaluate the CEO, which could be a huge deal for the nonprofit.
  - For example, if a nonprofit's mission is "Help animals everywhere," does this mean "Help as many animals as possible" (which might indicate a move toward focusing on farm animals) or "Help animals in the same way the nonprofit traditionally has" or something else? How does it imply the nonprofit should make tradeoffs between helping e.g. dogs, cats, elephants, chickens, fish or even insects? How a board member answers questions like this seems central to how their presence on the board is going to affect the nonprofit.
- They **need to have a personality and position capable of challenging the CEO** (though also capable of staying out of the way).
  - A common problem I see is that some board member is (a) not very engaged with the nonprofit itself, but (b) highly values their personal relationship with the CEO and other board members. This seems like a bad combination, but unfortunately a common one. Board members need to be willing and able to create conflict in order to do the right thing for the nonprofit.
  - Limiting the number of board members who are employees (reporting to the CEO) seems important for this reason.

- If you can't picture a board member "making waves," they probably shouldn't be on the board - that attitude will seem fine more than 90% of the time, but it won't work well in the rare cases where the board really matters.
- On the other hand, if someone is *only comfortable* "making waves" and feels useless and out of sorts when they're just nodding along, that person shouldn't be on the board either. As noted above, board members need to be ready for a weird job that involves stepping up when the situation requires it, but staying out of the way when it doesn't.
- They should probably have a **well-developed take on what their job is as a board member**. Board members who can't say much about where they expect to be highly engaged, vs. casually advisory - and how they expect to invest in getting the knowledge they need to do a good job leading on particular issues - don't seem like great bets to step up when they most need to (or stay out of the way when they should).

In my experience, most nonprofits are not looking for these qualities in board members. They are, instead, often looking for things like:

- Celebrity and reputation - board members who are generally impressive and well-regarded and make the nonprofit look good. Unfortunately, I think such people often just don't have much time or interest for their job. Many are also uninterested in causing any conflict, which makes them basically useless as board members IMO.
- Fundraising - a lot of nonprofits pretty much explicitly just try to put people on the board who will help raise money for them. This seems bad for governance.
- Narrow expertise on some topic that is important for the nonprofit. I don't really think this is what nonprofits should be seeking from board members,<sup>10</sup> except to the extent it ties deeply into the board members' core duties, e.g., where it's important to have an independent view on technical topic X in order to do a good job evaluating the CEO.

I think a good profile for a board member is someone who cares greatly about the nonprofit's mission, and wants it to succeed, to the point where they're ready to have tough conversations if they see the CEO falling short. Examples of such people might be major funders, or major stakeholders (e.g., a community leader from a community of people the nonprofit is trying to help).

## A few practices that seem good

I'll anticlimactically close with a few practices that seem helpful to me. These are mostly pretty generic practices, useful for both for-profit and nonprofit boards, that I have seen working in practice but also seen too many boards going without. They don't fully address the weirdnesses discussed above (especially the stuff specific to nonprofit as opposed to for-profit boards), but they seem to make things some amount better.

**Keeping it simple for low-stakes organizations.** If a nonprofit is a year old and has 3 employees, it probably shouldn't be investing a ton of its energy in having a great board (especially since this is hard).

A key question is: "If the board just stays checked out and doesn't hold the CEO accountable, what's the worst thing that can happen?" If the answer is something like "The nonprofit's relatively modest budget is badly spent," then it might not be worth a huge investment in building a great board (and in taking some of the measures listed

below). Early-stage nonprofits often have a board consisting of 2-3 people the founder trusts a lot (ideally in a "you'd fire me if it were the right thing to do" sense rather than in a "you've always got my back" sense), which seems fine. The rest of these ideas are for when the stakes are higher.

**Formal board-staff communication channels.** A very common problem I see is that:

- Board members know almost nothing about the organization, and so are hesitant to engage in much of anything.
- Employees of the organization know far more, but find the board members mysterious/unapproachable/scary, and don't share much information with them.

I've seen this dynamic improved some amount by things like a **staff liaison**: a board member who is designated with the duty, "Talk to employees a lot, offer them confidentiality as requested, try to build trust, and gather information about how things are going." Things like regular "office hours" and showing up to company events can help with this.

**Viewing board seats as limited.** It seems unlikely that a board should have more than 10 members (and even 10 seems like a lot), since it's hard to have a productive meeting past that point.<sup>11</sup> When considering a new addition to the board, I think the board should be asking something much closer to "Is this one of the 10 best people in the world to sit on this board?" than to "Is this person fine?"

**Regular CEO reviews.** Many nonprofits don't seem to have any formal, regular process for reviewing the CEO's performance; I think it's important to do this.

The most common format I've seen is something like: one board member interviews the CEO's direct reports, and perhaps some other people throughout the company, and integrates this with information about the organization's overall progress and accomplishments (often presented by the organization itself, but they might ask questions about it) to provide a report on what the CEO is doing well and could do better. I think this approach has a lot of limitations - staff are often hesitant to be forthcoming with a board member (even when promised anonymity), and the board member often lacks a lot of key information - but even with those issues, it tends to be a useful exercise.

**Closed sessions.** I think it's important for the board to have "closed sessions" where board members can talk frankly without the CEO, other employees, etc. hearing. I think a common mistake is to ask "Does anyone want the closed session today or can we skip it?" - this puts the onus on board members to say "Yes, I would like a closed session," which then implies they have something negative to say. I think it's better for whoever's running the meetings to identify logical closed sessions (e.g., "The board minus employees"), allocate time for them and force them to happen.

**Regular board reviews.** It seems like it would be a good idea for board members to regularly assess each other's performance, and the performance of the board as a whole. But I've actually seen very little of this done in practice and I can't point to versions of it that seem to have some track record of working well. It does seem like a good idea though!

## Conclusion

The board is the only body at a nonprofit that can hold the CEO accountable to accomplishing the mission. I broadly feel like most nonprofit boards just aren't very well-suited to this duty, or necessarily to much of anything. It's an inherently weird structure that seems difficult to make work.

I wish someone would do a great job studying and laying out how nonprofit boards should be assembled, how they should do their job and how they can be held accountable. You can think of this post as my quick, informal shot at that.









[Comment/discuss](#)

---

## Footnotes

1. I'm using the term "CEO" throughout, although the chief executive at a non profit sometimes has another title, such as "Executive Director." [↩](#)
2. A lot of this piece is about how the *fundamental setup* of a nonprofit board leads to the kinds of problems and dynamics I'm describing. This doesn't mean we should necessarily think there's any way to fix it or any better alternative. It just means that this setup seems to bring a lot of friction points and challenges that *most* relationships between supervisor-and-supervised don't seem to have, which

can make the experience of interacting with a board feel vaguely unlike what we're used to in other contexts, or "weird."

People who have interacted with tons of boards might get so used to these dynamics that they no longer feel weird. I haven't reached that point yet myself though.

- ↳
3. The fact that the nonprofit's goals aren't clearly defined and have no clear metric (and often aren't susceptible to measurement at all) is a pretty general challenge of nonprofits, but I think it especially shows up for a structure (the board) that is already weird in the various other ways I'm describing. ↳
  4. Superficially, you could make most of the same complaints about shareholders of a for-profit company. But:
    - Shareholders are the people who ultimately make or lose money if the company does well or poorly (you can think of this as a form of accountability). By contrast, nonprofit board members often have very little (or only an idiosyncratic) personal connection to and investment in the organization.
    - Shareholders compensate for their low engagement by picking representatives (a board) whom they can hold accountable for the company's performance. Nonprofit board members are the representatives, and aren't accountable to anyone. ↳
  5. Especially "good and concise." Most of the points I make here can be found in some writings on boards somewhere, but it's hard to find sensible-seeming and comprehensive discussions of what the board should be doing and who should be on it. ↳
  6. Part of the CEO's job is fundraising, and if they do a bad job of this, it's going to be obvious. But that's only part of the job. At a nonprofit, a CEO could easily be bringing in plenty of money and just doing a horrible job at the mission - and if the board isn't able to learn this and act on it, it seems like very bad news. ↳
  7. The charter and bylaws are like the "constitution" of a nonprofit, laying out how its governance works. ↳
  8. This is a judgment call, and one way to approach it would be to reserve something like 1 hour of full-board meeting time per year for talking about these sorts of things (and pouring in more time if at least, like, 1/3 of the board thinks something is a big deal).

Some examples of things I think are and aren't usually a big enough deal to start paying serious attention to:

- Big enough deal: financial decisions that increase the odds of going "belly-up" (running out of money and having to fold) by at least 10 percentage points. Not a big enough deal: spending money in ways that are arguably bad uses of money, having a lowish-but-not-too-far-off-of-peer-organizations amount of runway.
- Big enough deal: deficiencies in financial controls that an auditor is highlighting, or a lack of audit altogether, until a plan is agreed to to address these things. Not a big enough deal: most other stuff in this category.
- Big enough deal: organizations with substantial "PR risk" exposure should have a good team for assessing this and a "crisis plan" in case something happens. Not a big enough deal: specific organizational decisions and practices that you are not personally offended by or find unethical, but

could imagine a negative article about. (If you do find them substantively unethical, I think that's a big enough deal.)

- Big enough deal: transferring like 1/3 or more of valuable things the nonprofit has (intellectual property, money, etc.) to another entity not controlled by the board. Not a big enough deal: starting an affiliate organization primarily for taking donations in another country or something.
- Big enough deal: doubling or halving the workforce. Not a big enough deal: smaller hirings and firings.

↳

9. Sometimes the Board Chair is the CEO, and sometimes the Chair is an employee of the company who also sits on the board. In these cases, I think it's good for there to be a separate Lead Independent Director who is not employed by the company and is therefore exclusively representing the Board. They can help set agendas, lead meetings, and take responsibility by default when it's otherwise unclear who would do so. ↳
10. Nonprofits can get expertise on topic X by hiring experts on X to advise them. The question is: when is it important to have an expert on X *evaluating the CEO?* ↳
11. Though it could be fine and even interesting to have giant boards - 20 people, 50 or more - that have some sort of "executive committee" of 10 or fewer people doing basically all of the meetings and all of the work (with the rest functioning just as very passive, occasionally-voting equivalents of "shareholders"). Just assume I'm talking about the "executive committee" type thing here. ↳

# On A List of Lethalities

Response to (Eliezer Yudkowsky): [A List of Lethalities](#).

*Author's Note: I do not work in AI Safety, lack technical domain knowledge and in many ways am going to be wrong. I wasn't going to write this to avoid potentially wasting too much time all around without having enough to offer, and for fear of making stupid errors, but it was clear that many people thought my response would be valuable. I thank those whose anonymous sponsorship of this post both paid for my time and made me update that the post was worth writing. I would be happy for this to happen again in the future.*

Eliezer has at long last delivered the definitive list of Eliezer Rants About Why AGI Will Definitely Absolutely For Sure Kill Everyone Unless Something Very Unexpected Happens.

This is excellent. In the past we had to make do with makeshift scattershot collections of rants. Now they are all in one place, with a helpful classification system. Key claims are in bold. We can refer, consider and discuss them.

It would be an even better post if it were more logically organized, with dependencies pointed out and mapped and so on.

One could also propose making it not full of rants, but I *don't* think that would be an improvement. The rants are *important*. The rants contain *data*. They reveal Eliezer's cognitive state and his assessment of the state of play. *Not* ranting would leave important bits out and give a meaningfully misleading impression.

I am reminded of [this comment of mine](#) that I dug out of the archives, [on another Eliezer post](#) that was both useful and enthused with this kind of attitude:

Things I instinctively observed slash that my model believes that I got while reading that seem relevant, not attempting to justify them at this time:

1. There is a core thing that Eliezer is trying to communicate. It's not actually about timeline estimates, that's an *output* of the thing. Its core message length is short, but all attempts to find short ways of expressing it, so far, have failed.
2. Mostly so have *very long* attempts to communicate it and its prerequisites, which to some extent at least includes the Sequences. Partial success in some cases, full success in almost none.
3. This post, and this whole series of posts, feels like its primary function is *training data* to use to produce an Inner Eliezer that has access to the core thing, or even better to know the core thing in a fully integrated way. And maybe a lot of Eliezer's *other* communications is kind of also trying to be similar training data, no matter the superficial domain it is in or how deliberate that is.
4. The condescension is *important information* to help a reader figure out what is producing the outputs, and hiding it would make the task of 'extract the key insights' harder.
5. Similarly, the repetition of the same points is also potentially important information that points towards the core message.
6. That doesn't mean all that isn't super annoying to read and deal with, especially when he's telling *you in particular* that you're wrong. Cause it's totally that.
7. There are those for whom this makes it easier to read, especially given it is *very long*, and I notice both effects.
8. My Inner Eliezer says that writing this post without the condescension, or making it shorter, would be *much much* more effort for Eliezer to write. To the extent such a thing can be written, someone else has to write that version. Also, it's kind of text in several places.
9. The core message is what matters and the rest mostly doesn't?
10. I am arrogant enough to think I have a non-zero chance that I know enough of the core thing and have enough skill that with enough work I could perhaps find an improved way to communicate it given the new training data, and I have the urge to try this impossible-level problem if I could find the time and focus (and help) to make a serious attempt.

[Reply](#)

Most of this applies again. Eliezer says explicitly that the alternative post would have been orders of magnitude harder to write, and that the attitude is important information.

I would expand this. *Not only* are the attitude and repetition important information in terms of allowing you to understand the algorithm generating the post and create a better Inner Eliezer, but they *also* are importantly illustrating *the cognitive world in which Eliezer is operating*.

The fact that *this is the post we got*, as opposed to a different (in many ways better) post, is a reflection of the fact that our Earth is failing to understand what we are facing. It is failing to look the problem in the eye, let alone make real attempts at solutions.

Eliezer is not merely talking to *you, yes you* (with notably rare exceptions) when he does this. He is also saying *model the world as if it really is forcing him to talk like this*.

The only point above that doesn't seem to apply here is #9.

The core message remains the most important thing. Conveying the core message alone would be a big win. But here it *also* matters that people grasp as many of the individual

points as possible, especially whichever of them happens to be the one bottlenecking their understanding of the scope and difficulty of the problem or allowing them to rationalize.

Thus there needs to be a second version of the document that *someone else* writes that contains the properly organized details without the ranting, for when that is what is needed.

In terms of timelines, only ‘endgame’ timelines (where endgame means roughly ‘once the first team gets the ability to create an AGI capable of world destruction’) are mentioned in this post, because they are a key part of the difficulty and ‘how long it takes to get there’ mostly isn’t. Talk of *when* AGI will kill us is distinct from talk of *how* or *why* it will, or *whether* it will be built. That stuff was the subject of that other post, and it *doesn’t really matter* in this context.

It is central to the doom claim that once one group can build an AGI, other groups also rapidly gain this ability. This forces humanity to solve the problem both *on the first try* and also *quickly*, a combination that makes an otherwise highly difficult but potentially solvable problem all but impossible. I find this plausible but am in no way confident in it.

I will also be assuming as a starting point the ability of at least one group somewhere to construct an AGI on some unspecified time frame.

## Goals

The goal of the bulk of the post is both to give my reactions to the individual claims and to attempt to organize them into a cohesive whole, and to see where my model differs from Eliezer’s even after I get access to his.

Rather than put the resulting summary results at the bottom, I’m going to put them at the top where they’ll actually get read, then share my individual reasoning afterwards because actually reasoning this stuff out loud seems like The Way.

## Summary of List, Agreements and Disagreements

Some of what the post is doing is saying ‘here is a particular thing people say that is stupid and wrong but that people use as an excuse, and here is the particular thing I say in response to that.’ I affirm these one by one below.

More centrally, the post is generated by a very consistent model of the situation, so having thought about each individual statement a summary here is more like an attempt to recreate the model generating the points rather than the points themselves.

To the extent that I am *wrong* about the contents of the generative model, that seems important to clarify.

I would say my takeaways are here, noting they are in a different order than where they appear in the post:

M1. Creating a powerful unsafe AGI quickly kills everyone. No second chances.

M2. The only known pivotal acts that stop the creation of additional powerful AGIs all require a powerful AGI. Weak systems won’t get it done.

M3. AGI will happen mostly on schedule unless stopped by such a pivotal act, whether or not it is safe. So not only do we only get one chance to solve the problem of alignment, we don’t get much time. Within two years of the first group’s ability to build an (unsafe) AGI, five more groups can do so including Facebook. Whoops.

M4. Powerful AGI is dramatically different and safety strategies that work on weak AGIs won't work on powerful ones.

M5. Most safety ideas and most safety work are known to be useless and have no value in terms of creating safe powerful AGIs. All the usual suspects don't work for reasons that are listed, and there are many reasons the problem is extremely difficult.

M6. We have no plan for how to do anything useful. No one who isn't Eliezer seems capable of even understanding the problems well enough to explain them, and no one who can't explain the problems is capable of nontrivially useful AI Safety work.

M7 (*not explicitly said but follows and seems centrally important*). Most attempts to create AI Safety instead end up creating AI capability work, and the entire attempt has so far been net negative, and is likely net negative even if you exclude certain large obviously negative projects.

M8. We have no idea what the hell is going on with these systems. Even if we did, that would break down once we started using observations while training AIs.

M9. The problem would *still* be solvable if a failed attempt didn't kill everyone and we had enough time. We get neither. Attempts that can't kill you aren't real attempts and don't tell you if your solution works.

M10 (*let's just say it*). Therefore, DOOM.

That is my summary. As Eliezer notes, different people will need to hear or learn different parts of this, and would write different summaries.

Based on this summary, which parts do I agree with? Where am I skeptical?

For all practical purposes I *fully agree* with M1, M4, M5, M7 (!) and M9.

For all practical purposes I *mostly agree* with M2, M6 and M8, but am less confident that the situations are as extreme as described.

For M2 I hold out hope that an as-yet-unfound path could be found.

For M6 I do not think we can be so confident there aren't valuable others out there (although obviously not as many as we need/want).

For M8, I do not feel I am in a position to evaluate our future ability to look inside the inscrutable matrixes enough to have so little hope.

For M10, I agree that M10 follows from the M1-M9, and unconditionally agree that there is a highly unacceptable probability of doom even if all my optimistic doubts are right.

I am least convinced of M3.

M3 matters a lot. M3 is stated most directly in Eliezer's #4, where a proof is sketched:

**#4. We can't just "decide not to build AGI"** because GPUs are everywhere, and knowledge of algorithms is constantly being improved and published; 2 years after the leading actor has the capability to destroy the world, 5 other actors will have the capability to destroy the world.

In particular, I question the assumption that incremental improvement in the knowledge of algorithms and access to GPUs is sure to be sufficient to generate AGI, or that there is no plausible hard step or secret sauce that could buy you a substantial lead without being published or stolen immediately in a way that invalidated that lead, and that there is no

possibility of a flat out ‘competence gap’ or capacity gap of some kind that matters, and that essentially unlimited numbers of additional efforts will necessarily be close behind.

This also seems closely related to #22’s claim that there is a simple core to general intelligence, which I am also not yet convinced about.

Thus, I am neither convinced that doom is coming especially quickly, nor that it will involve an AGI that looks so much like our current AIs, nor am I convinced that the endgame window will be as short as the post assumes.

I do agree that this scenario is *possible*, and has non-trivial probability mass. That is more than enough to make the current situation unacceptable, but it is important to note where one is and is not yet convinced.

I do agree that you likely don’t *know* how much time you have, even if you think you may have more time.

I strongly agree that creating an aligned AI is harder, probably much harder, than creating an unaligned AI, that it requires additional work and additional time if it can be done at all, and that if it needs to be done both quickly and without retries chances of success seem extremely low.

I have a lot of other questions, uncertainties, brainstorms and disagreements in the detail section below, but those are the ones that *matter* for the core conclusions and implications.

Even if those ‘optimistic doubts’ proved true, *mostly* it doesn’t change what needs to be done or give us an idea of how to do it.

## Preamble

-3: Yes, both the [orthogonality thesis](#) and [instrumental convergence](#) are true.

-2: When we say Alignment at this point we mean something that can carry out a pivotal task that prevents the creation of another AGI while having less than a 50% chance of killing a billion people. Anything short of mass death, and we’ll take it.

-1: The problem is so difficult because we need to solve the problem on the first critical try on a highly limited time budget. The way humans typically solve hard problems involves taking time and failing a lot, which here would leave us very dead. If we had time (say 100 years) and unlimited retries the problem is still super hard but (probably?) eminently solvable by ordinary human efforts.

## Section A

1. AGI will not be upper-bounded by human ability or human learning speed. Things much smarter than human would be able to learn from less evidence than humans require.

...

It is not naturally (by default, barring intervention) the case that everything takes place on a timescale that makes it easy for us to react.

Yes, obviously.

This is a remarkably soft-pedaling rant. Given sufficient processing power, anything the AGI can learn from what data it has is something it already knows. Any skill it can develop is a skill it already has.

2. A cognitive system with sufficiently high cognitive powers, given any medium-bandwidth channel of causal influence, will not find it difficult to bootstrap to overpowering capabilities independent of human infrastructure.

...

Losing a conflict with a high-powered cognitive system looks at least as deadly as "everybody on the face of the Earth suddenly falls over dead within the same second".

Yes, obviously.

If you don't like the nanotech example (as some don't), ignore it. It's not important. A sufficiently intelligent system that is on the internet or can speak to humans simply wins, period. The question is what counts as sufficiently intelligent, not whether there is a way.

3. We need to get alignment right on the 'first critical try' at operating at a 'dangerous' level of intelligence, where unaligned operation at a dangerous level of intelligence kills everybody on Earth and then we don't get to try again.

Yes, obviously this is the default outcome.

If it's smart enough to figure out how to do things that prevent other AGIs it is also almost certainly smart enough to figure out how to kill us and by default that is going to happen because it makes it easier to achieve the AGI's goals whatever they are.

I can see arguments for why the chance you get a second shot is not zero, but it is very low.

4. We can't just "decide not to build AGI" because GPUs are everywhere, and knowledge of algorithms is constantly being improved and published; 2 years after the leading actor has the capability to destroy the world, 5 other actors will have the capability to destroy the world.

This is NOT obvious to me.

This is making assumptions about what physically results in AGI and how information develops and spreads. I notice I don't share those assumptions.

It seems like this is saying either that there are no 'deep insights' left before AGI, or that any such deep insights will either (A) inevitably happen in multiple places one after another or (B) will inevitably leak out quickly in a form that can be utilized.

It also says that there won't be a big 'competence gap' between the most competent/advanced group and 6th such group, so within 2 years the others will have caught up. That there won't be any kind of tacit knowledge or team skill or gap in resources or willingness to simply do the kind of thing in question at the sufficient level of scale, or what have you.

I do not see why this should be expected with confidence.

Yes, we have seen AI situations in which multiple groups were working on the same problem, most recently image generation from a text prompt, and finished in similar time frames. It can happen, especially for incremental abilities that are mostly about who feels like spending compute and manpower on improving at a particular problem this year instead of last year or next year. And yes, we have plenty of situations in which multiple start-ups were racing for a new market, or multiple scientists were racing for some discovery, or whatnot.

We also have plenty of situations in which there was something that could have been figured out at any time, and it just kind of wasn't for quite a while. Or where something was being done quite stupidly and badly for a very long time. Or where someone figured something

out, tried to tell everyone about their innovation, and everyone both ignored them and didn't figure it out on their own for a very long time.

Certainly a substantial general capacity advantage, or a capacity advantage in the place that turns out to matter, seems highly plausible to me.

From his other writings it is clear that a lot of this is Eliezer's counting on the code being stolen and that it will be possible to remove whatever safeties are in place. I agree with the need for real security to prevent this when the time comes and the worry that scale may make such security unrealistic and expensive, but also this assumes a kind of competence from the people knowing to steal the code, and also a competence that they can use what they steal, whereas I'm done assuming such competencies will exist at all.

I'm not saying the baseline scenario here is *impossible* or even all that unlikely, but it seems quite possible for it not to be the case, or at least for the numbers quoted above to not be.

That doesn't solve the problem of the underlying dynamic. There is still *some* time limit. Even if there is a good chance that you can indeed 'decide not to build AGI' for a while, there is still a continuous risk that you are wrong about that, and there are still internal pressures not to wait for other reasons, and all that.

**5. We can't just build a very weak system**, which is less dangerous because it is so weak, and declare victory; because later there will be more actors that have the capability to build a stronger system and one of them will do so. I've also in the past called this the 'safe-but-useless' tradeoff, or 'safe-vs-useful'. People keep on going "why don't we only use AIs to do X, that seems safe" and the answer is almost always either "doing X in fact takes very powerful cognition that is not passively safe" or, even more commonly, "because restricting yourself to doing X will not prevent Facebook AI Research from destroying the world six months later".

Fundamentally, yes. You either do a pivotal act that stops other AGIs from being constructed or you don't. Doing one requires non-safe cognition. Not doing one means someone else creates non-safe cognition. No good.

**6. We need to align the performance of some large task, a 'pivotal act' that prevents other people from building an unaligned AGI that destroys the world.** While the number of actors with AGI is few or one, they must execute some "pivotal act", strong enough to flip the gameboard, using an AGI powerful enough to do that. It's not enough to be able to align a *weak system* - we need to align a system that can do some single *very large thing*. The example I usually give is "burn all GPUs".

...

Yes. I notice I skipped ahead to this a few times already. I probably would have moved the order around.

It takes a lot of power to do something to the current world that prevents any other AGI from coming into existence; nothing which can do that is passively safe in virtue of its weakness.

## **7. There are no pivotal weak acts.**

I am not as convinced that there don't exist pivotal acts that are importantly easier than directly burning all GPUs (after which I might or might not then burn most of the GPUs anyway). There's no particular reason *humans* can't perform dangerous cognition *without* AGI help and do some pivotal act on their own, our cognition is not exactly safe. But if I did have such an idea that I thought would work I wouldn't write about it, and it most certainly wouldn't be in the Overton window. Thus, I do not consider the failure of our public discourse to generate such an act to be especially strong evidence that no such act exists.

8. The best and easiest-found-by-optimization algorithms for solving problems we want an AI to solve, readily generalize to problems we'd rather the AI not solve

Yes, obviously.

9. The builders of a safe system, by hypothesis on such a thing being possible, would need to operate their system in a regime where it has the *capability* to kill everybody or make itself even more dangerous, but has been successfully designed to not do that. **Running AGIs doing something pivotal are not passively safe**, they're the equivalent of nuclear cores that require actively maintained design properties to not go supercritical and melt down.

Yes, obviously, for the combined human-AI system doing the pivotal thing. Again, one can imagine putting all the unsafe cognition 'into the humans' in some sense.

## Section B.1

10. On anything like the standard ML paradigm, you would need to somehow generalize optimization-for-alignment you did in safe conditions, across a big distributional shift to dangerous conditions.

...

Powerful AGIs doing dangerous things that will kill you if misaligned, must have an alignment property that generalized far out-of-distribution from safer building/training operations that didn't kill you. This is where a huge amount of lethality comes from on anything remotely resembling the present paradigm.

...

10a. Note that anything substantially smarter than you poses a threat given *any* realistic level of capability. Eg, "being able to produce outputs that humans look at" is probably sufficient for a generally much-smarter-than-human AGI to navigate its way out of the causal systems that are humans, especially in the real world where somebody trained the system on terabytes of Internet text, rather than somehow keeping it ignorant of the latent causes of its source code and training environments.

Yes. 10 seems transparently and obviously true, yet it does need to be said explicitly.

I am labeling 10a because I consider it an important sub-claim, one that I am *highly* confident is true. A much-smarter-than-human AGI capable of getting its text read by humans will be able to get those humans to do what it wants, period. This is one of those no-it-does-not-seem-wise-to-explain-why-I-am-so-confident-this-is-true situations so I won't, but I am, again, very confident.

11. There is no pivotal act this weak; **there's no known case where you can entrain a safe level of ability on a safe environment where you can cheaply do millions of runs, and deploy that capability to save the world** and prevent the next AGI project up from destroying the world two years later. Pivotal weak acts like this aren't known, and not for want of people looking for them.

...

You don't get 1000 failed tries at burning all GPUs – because people will notice, even leaving out the consequences of capabilities success and alignment failure.

There certainly isn't a *publicly known* such act that could possibly be implemented, and there has definitely been a lot of public searching for one. It doesn't seem impossible that an answer exists and that those who find it don't say anything for very good reasons. Or that 'a

lot of trying to do X and failing' is surprisingly weak evidence that X is impossible, because the efforts are correlated in terms of their blind spots.

**12. Operating at a highly intelligent level is a drastic shift in distribution from operating at a less intelligent level**, opening up new external options, and probably opening up even more new internal choices and modes. Problems that materialize at high intelligence and danger levels may fail to show up at safe lower levels of intelligence, or may recur after being suppressed by a first patch.

Yes, yes, we said that already.

**13. Many alignment problems of superintelligence will not naturally appear at pre-dangerous, passively-safe levels of capability.** Consider the internal behavior 'change your outer behavior to deliberately look more aligned and deceive the programmers, operators, and possibly any loss functions optimizing over you'. This problem is one that will appear at the superintelligent level; if, being otherwise ignorant, we guess that it is among the *median* such problems in terms of how *early* it naturally appears in earlier systems, then around *half* of the alignment problems of superintelligence will first naturally materialize *after* that one first starts to appear.

On the headline statement, yes, yes, again, didn't we say that already?

The example is definitely a danger at the superhuman level, but it seems like it is also a danger at the human level. [Have... you met humans?](#) Also have you met dogs and cats, definitely sub-human intelligences? This is not an especially 'advanced' trick.

This makes sense, because figuring out that a problem that *doesn't* exist at human levels *will* exist at superhuman levels seems difficult by virtue of the people thinking about the problem being humans. We can figure out things that current systems maybe aren't doing, like 'pretend to be aligned to fool creators' because *we are intelligent systems that do these things*. And that seems like a problem it would be *very easy* to get to materialize early, in an actually safe system, because again existence proof and also it seems obvious *how* to do it. That doesn't mean I know *how* to solve the problem, but I can make it show up.

What are the problems that *don't* show up in sub-human AI systems and also don't show up in humans because we can't think of them? I don't know. I can't think of them. That's why they don't show up.

Thus, to the extent that we can talk about there being distinct alignment problems like this that one can try to anticipate and solve, the nasty ones that only show up in the one-shot final exam are going to be things that *we are not smart enough to think of* and thus we can't prepare for them. Which means we need a *general* solution, or else we're hoping there are no such additional problems.

**14. Some problems**, like 'the AGI has an option that (looks to it like) it could successfully kill and replace the programmers to fully optimize over its environment', **seem like their natural order of appearance could be that they first appear only in fully dangerous domains.**

...

Trying to train by gradient descent against that behavior, in that toy domain, is something I'd expect to produce not-particularly-coherent local patches to thought processes, which would break with near-certainty inside a superintelligence generalizing far outside the training distribution and thinking very different thoughts. Also, programmers and operators themselves, who are used to operating in not-fully-dangerous domains, are operating out-of-distribution when they enter into dangerous ones; our methodologies may at that time break.

Being able to somehow take control and override the programmers to take control of the reward function is, again, something that humans essentially do all the time. It is coming. The question is will fixing it in a relatively safe situation lead to a general solution to the problem?

My presumption is that if someone goes in with the goal of ‘get this system to stop having the problem’ the solution found has almost zero chance of working in the dangerous domain. If your goal is to actually figure out what’s going on in a way that might survive, then maybe there’s *some* chance? Still does not seem great. The thing we look to prevent may not meaningfully interact with the thing that is coming, at all.

**15. Fast capability gains seem likely, and may break lots of previous alignment-required invariants simultaneously.** Given otherwise insufficient foresight by the operators, I’d expect a lot of those problems to appear approximately simultaneously after a sharp capability gain. See, again, the case of human intelligence.

Yes.

When I said ‘yes’ above I wasn’t *at all* relying on the example of human intelligence, or the details described later, but I’m going to quote it in full because this is the first time it seems like an especially valuable detailed explanation.

We didn’t break alignment with the ‘inclusive reproductive fitness’ outer loss function, immediately after the introduction of farming – something like 40,000 years into a 50,000 year Cro-Magnon takeoff, as was itself running very quickly relative to the outer optimization loop of natural selection. Instead, we got a lot of technology more advanced than was in the ancestral environment, including contraception, in one very fast burst relative to the speed of the outer optimization loop, late in the general intelligence game. We started reflecting on ourselves a lot more, started being programmed a lot more by cultural evolution, and lots and lots of assumptions underlying our alignment in the ancestral training environment broke simultaneously.

(People will perhaps rationalize reasons why this abstract description doesn’t carry over to gradient descent; eg, “gradient descent has less of an information bottleneck”. My model of this variety of reader has an inside view, which they will label an outside view, that assigns great relevance to some other data points that are *not* observed cases of an outer optimization loop producing an inner general intelligence, and assigns little importance to our one data point actually featuring the phenomenon in question. When an outer optimization loop actually produced general intelligence, it broke alignment after it turned general, and did so relatively late in the game of that general intelligence accumulating capability and knowledge, almost immediately before it turned ‘lethally’ dangerous relative to the outer optimization loop of natural selection. Consider skepticism, if someone is ignoring this one warning, especially if they are not presenting equally lethal and dangerous things that they say will go wrong instead.)

I both agree that the one data point is not being given enough respect, and also don’t think you need the data point. There are going to be a whole lot of things that are true about a system when the system is insufficiently intelligent/powerful that won’t be true when the system gets a lot more intelligent/powerful and some of them are things you did not realize you were relying upon. It’s going to be a problem.

## Section B.2

16. Even if you train really hard on an exact loss function, that doesn’t thereby create an explicit internal representation of the loss function inside an AI that then continues to pursue that exact loss function in distribution-shifted environments

...

outer optimization even on a very exact, very simple loss function doesn't produce inner optimization in that direction.

This happens *in practice in real life*, it is what happened in *the only case we know about*, and it seems to me that there are deep theoretical reasons to expect it to happen again

Yes. It won't do that, not if your strategy is purely to train on the loss function. There is no reason to expect it to happen. So don't do that. Need to do something else.

17. In the current optimization paradigm there is no general idea of how to get particular inner properties into a system, or verify that they're there, rather than just observable outer ones you can run a loss function over.

I think we have *some* ability to verify if they are there? As in, Chris Olah and a few others have made enough progress that at least some current-paradigm systems for which they can identify some of the inner properties of the system, with expectation of more in the future. They have no idea how to choose or cause those properties that I know about, but there's at least some hope for some observability.

If you can observe it, you can at least in theory train on it as well, although that risks training the AI to make your observation method stop working? As in, suppose you have a classifier program. From my conversations, it sounds like at least sometimes you can say 'this node represents whether there is a curve here' or whatever. If you can do that, presumably (at least in theory) you can then train or do some sort of selection on whether or not that sort of thing is present and in what form, and iterate, and you can have at least some say over how the thing you eventually get is structured within the range of things that could possibly emerge from your loss function, or something. There are other things I can think of to try as well, which of course are probably obvious nonsense, or worse nonsense just non-obvious enough to get us all killed, but you never know.

18. There's no reliable Cartesian-sensory ground truth (reliable loss-function-calculator) about whether an output is 'aligned', because some outputs destroy (or fool) the human operators and produce a different environmental causal chain behind the externally-registered loss function.

Yes, that is a thing. You are in fact hoping that it importantly *doesn't* optimize too well for what reward signal it gets and instead optimizes on your intent. That seems hard.

**19.** More generally, **there is no known way to use the paradigm of loss functions, sensory inputs, and/or reward inputs, to optimize anything within a cognitive system to point at particular things within the environment** – to point to *latent events and objects and properties in the environment*, rather than *relatively shallow functions of the sense data and reward*.

Yes, I did realize that you'd said this already, but also it's seeming increasingly weird and like something you can overcome? As in, sure, you'll need to do something innovative to make this work and it's important to note that a lot of work has been done and no one's done it yet and that is quite a bad sign, but... still?

**20.** Human operators are fallible, breakable, and manipulable. **Human raters make systematic errors – regular, compactly describable, predictable errors.** To *faithfully* learn a function from 'human feedback' is to learn (from our external standpoint) an unfaithful description of human preferences, with errors that are not random (from the outside standpoint of what we'd hoped to transfer). If you perfectly learn and perfectly maximize *the referent of rewards assigned by human operators*, that kills them. It's a fact about the territory, not the map – about the environment, not the optimizer – that the *best predictive explanation* for human answers is one that predicts the systematic errors in our responses, and therefore is a psychological concept that

correctly predicts the higher scores that would be assigned to human-error-producing cases.

I worry that there's a leap in here and it's taking the principle of 'almost every possible AGI kills you' too far. In general, I am *totally on board* with the principle that almost every possible AGI kills you. Most of the time that the post says 'so it kills you' this is definitely the thing that happens next if the previous things did indeed take place.

If by 'fool the operators' we mean things like 'take control of the operators and implant a chip in their head' then yes, there is that, but that doesn't seem like what is being described here. What is being described here is your friendly neighborhood AGI that wants you to like its output, to really like it, so it tells you what you will be happy to hear every time even if the results would be quite bad.

Does that kill you (as in, kill everyone)?

It certainly *could* kill you. Certainly it will intentionally choose errors over correct answers in some situations. But so will humans. So will politicians. We don't exactly make the best possible decisions or avoid bias in our big choices. This seems like a level of error that is often going to be survivable. It depends on how the humans rely on it and if the humans know to avoid situations in which this will get them killed.

I believe that if you gave Eliezer or myself the job of using an AGI that was aligned *exactly* to the evaluations of its output by a realistically assembled team of human evaluators *on an individual answer basis*, as in it wasn't trained to play a long game to get stronger future evaluations and was merely responding to human bias, that this would be good enough for Eliezer's threshold of alignment – we would be a favorite to successfully execute a pivotal act without killing a billion or more people.

That doesn't mean this isn't a problem. This is *much worse* a scenario than if the AGI was somehow magically aligned to what we *should* in some sense rate its output, and this is going to compound with other problems, but solving *every problem except this one* does seem like it would bring us home.

There's something like a single answer, or a single bucket of answers, for questions like 'What's the environment really like?' and 'How do I figure out the environment?' and 'Which of my possible outputs interact with reality in a way that causes reality to have certain properties?', where a simple outer optimization loop will straightforwardly shove optimizees into this bucket.

When you have a wrong belief, reality hits back at your wrong predictions. When you have a broken belief-updater, reality hits back at your broken predictive mechanism via predictive losses, and a gradient descent update fixes the problem in a simple way that can easily cohere with all the other predictive stuff.

In contrast, when it comes to a choice of utility function, there are unbounded degrees of freedom and multiple reflectively coherent fixpoints. Reality doesn't 'hit back' against things that are locally aligned with the loss function on a particular range of test cases, but globally misaligned on a wider range of test cases.

...

## 21. The central result: **Capabilities generalize further than alignment once capabilities start to generalize far.**

Yes, although not obviously. The explanation in this bullet point is very non-intuitive to me. That's assuming I actually grok it correctly, which I *think* I did after reflection but I'm not sure. It's certainly not how I would think about or explain the conclusion at all, nor am I convinced the reasoning steps are right.

When you have a wrong belief that causes wrong predictions, you might or might not end up with a loss function that needs correction. It happens if the wrong predictions are inside the training set (or ancestral environment) and also have consequences that impact your loss function, which not all errors do. The argument is some combination of (A) that optimizing for local capabilities is more inclined to produce a generalizable solution than optimizing for local alignment, and (B) that you are likely to get alignment ‘wrong’ via aligning to a proxy measure in a way that will prove very wrong outside the training set and get you killed and will be in a utility function that will be fixed in place, whereas the capabilities can continue to adjust and improve in addition to your proxy measures being less likely to break.

Both arguments do seem largely right, or at least likely enough to be right that we should presume they are probably right in practice when it counts.

**22.** There’s a relatively simple core structure that explains why complicated cognitive machines work; which is why such a thing as general intelligence exists and not just a lot of unrelated special-purpose solutions; which is why capabilities generalize after outer optimization infuses them into something that has been optimized enough to become a powerful inner optimizer. The fact that this core structure is simple and relates generically to [low-entropy high-structure environments](#) is why humans can walk on the Moon. **There is no analogous truth about there being a simple core of alignment,** especially not one that is even easier for gradient descent to find than it would have been for natural selection to just find ‘want inclusive reproductive fitness’ as a well-generalizing solution within ancestral humans. Therefore, capabilities generalize further out-of-distribution than alignment, once they start to generalize at all.

Probably, but seems overconfident. Certainly natural selection did not find one, but that is far from an impossibility proof. General intelligence turned out to be, in a broad sense, something that could be hill climbed towards, which wasn’t true for some sort of stricter alignment. Or at least, it is not true yet. This is one of those problems that seems like it kind of *didn’t come up* for natural selection until quite recently.

A simple general core alignment, that fixes things properly in place in a way that matters, could easily have been *quite the large handicap* over time until very recently by destroying degrees of freedom.

The same way that we don’t need to align our current weaker AIs in ways that would be relevant to aligning strong AIs, nor would there have been much direct benefit to doing so, the same seems like it should hold true for everything made by natural selection until humans, presumably until civilization, and plausibly until industrial civilization or even later than that. At what point were people ‘smart enough’ in some sense, with enough possible out-of-sample plays, where ‘want inclusive reproductive fitness’ as an explicit goal would have started to outcompete the alternatives rather than some of that being part of some sort of equilibrium situation?

(I mean, yes, we *do* need to align *current* AIs (that aren’t AGIs) operating in the real world and our failure to do so is causing major damage *now*, but again at least this is a case of it being bad but not killing us yet.)

It took natural selection quite a long time in some sense to find general intelligence. How many cycles has it had to figure out a simple core of alignment, provided one exists?

We don’t know about a simple core of alignment. One might well not exist even in theory, and it would be good for our plan not to be counting on finding one. Still, one might be out there to be found. Certainly one *on the level of complexity of general intelligence* seems plausibly out there to be found slash seems highly likely to *not have already been found by natural selection* if it existed, and I don’t feel our current level of work on the problem is conclusive either – it’s more like there are all these impossible problems it has to solve, which are all the other points, and that’s the primary reason to be pessimistic about this.

**23. Corrigibility is anti-natural to consequentialist reasoning;** “you can’t bring the coffee if you’re dead” for almost every kind of coffee. We (MIRI) [tried and failed](#) to find a coherent formula for an agent that would let itself be shut down (without that agent actively trying to get shut down). Furthermore, many anti-corrigible lines of reasoning like this may only first appear at high levels of intelligence.

Yes. I too have found this to be one of the highly frustrating things to watch people often choose not to understand, or pretend not to understand (or, occasionally, actually not understand).

Corrigibility really, really isn’t natural, it’s super weird, it very much does not want to happen. This problem is very hard, and failing to solve it makes all the other problems harder.

I want to emphasize here, like in a few other places, that 99%+ of all people need to take in the message ‘corrigibility is anti-natural and stupidly hard’ rather than the other way around.

However, I am in sharing my thoughts and reactions and models mode, and while 99% of people need to hear one thing the remaining people end up being rather important, so: while not fooling myself in any way that this isn’t close to impossible, the good news is that I still kind of see this as something that is *less impossible* than some other impossible things, especially if we follow the highly useful ‘in the one case we know about’ principle and look at humans, we *do* see some humans who are functionally kind of corrigible in the ways that matter here, and I don’t *think* it involves having those humans believe a false thing (I mean they do, all humans do anyway, which could be doing a lot of the work, but that doesn’t seem like the central tech here).

The technology (in humans) is that the human values *the continued well-functioning of the procedure that generates the decision whether to shut them down* more than they care about whether the shut down occurs in worlds where they are shut down. Perhaps because the fact that the humans are shutting them down is evidence that they should be shut down, whereas engineering the humans to shut them down wouldn’t provide that evidence.

They will still do things within the rules of the procedure to convince you not to shut them down, but if you manage to shut them down anyway, they will abide by that decision. And they will highly value passing this feature on to others.

This corrigibility usually has its limits, in particular it breaks down when you talk about making the human *dead* or otherwise causing them to expect sufficiently dire consequences, either locally or globally.

Is the Constitution a suicide pact? It wouldn’t work if it wasn’t willing to be a *little bit* a suicide pact. It’s also obviously not fully working in the sense that it isn’t a suicide pact, and almost no one has any intention of letting it become one in a sufficiently obvious pinch. As a fictional and therefore clean example, consider the movie [Black Panther](#) – should you let yourself be challenged and shut down in this spot, given the consequences, because the rules are the rules, despite the person you’re putting in charge of those rules clearly having no inclination to care about those rules?

Thus, the utility function that combines ‘the system continuing to persevere is super important’ with the desire for other good outcomes is, under the hood, profoundly weird and rather incoherent, and very anti-natural to consequentialist reasoning. I have no doubt that the current methods would break down if tried in an AGI.

Which makes me wonder the extent to which the consequentialist reasoning is going too far and thus part of the problem that needs to be solved, but I don’t see how to get us out of this one yet, even in theory, without making things much worse.

In any case, I'm sure that is all super duper amateur hour compared to the infinite hours MIRI spent on this particular problem, so while I'm continuing my pattern of not giving up on the problem or declaring it unsolvable it is almost certainly not easy.

**24.** There are two fundamentally different approaches you can potentially take to alignment, which are unsolvable for two different sets of reasons; therefore, **by becoming confused and ambiguating between the two approaches, you can confuse yourself about whether alignment is necessarily difficult.**

The first approach is to build a CEV-style Sovereign which wants exactly what we extrapolated-want and is therefore safe to let optimize all the future galaxies without it accepting any human input trying to stop it.

The second course is to build corrigible AGI which doesn't want exactly what we want, and yet somehow fails to kill us and take over the galaxies despite that being a convergent incentive there.

I am basically a CEV skeptic, in the sense that my model of Eliezer thinks it is impossible to implement on the first try but if you did somehow implement it then it would work. Whereas I think that not only is the problem impossible but also if you solved the impossible problem I am predicting a zero-expected-value outcome *anyway*. I don't even think the impossible thing works *in theory*, at least as currently theorized.

Whereas I'm a mild corrigibility optimist in the sense that I do recognize it's an impossible problem but it does at least seem like a relatively solvable impossible problem even if attempts so far have not gotten anywhere.

I'm also not convinced that the get-it-right-on-first-try approach has to go through CEV, but details there are both beyond scope of the question here and also I'm likely very out of my depth, so I'll leave that at that.

I haven't experienced that much frustration on *this particular* dilemma, where people don't know if they're trying to get things right on the first try or they're trying to solve corrigibility, but that's probably because I've never fully been 'in the game' on this stuff, so I consider that a blessing. I do not doubt the reports of these ambiguations.

## Section B.3

**25. We've got no idea what's actually going on inside the giant inscrutable matrices and tensors of floating-point numbers.** Drawing interesting graphs of where a transformer layer is focusing attention doesn't help if the question that needs answering is "So was it planning how to kill us or not?"

Yes, at least for now this is my understanding as well.

I have never attempted to look inside a giant inscrutable matrix. Even if we did have *some* idea what is going on inside in some ways, that does not tell us whether the machine is trying to kill us. And if we could look inside and tell, all we'd be doing is teaching the machine to figure out how to hide from our measurements that it was trying to kill us, or whatever else it was up to that we didn't like, including hiding that it was hiding anything. So there's that.

I have heard claims that interpretability is making progress, that we have *some* idea about *some* giant otherwise inscrutable matrices and that this knowledge is improving over time. I do not have the bandwidth that would be required to evaluate those claims and I don't know how much usefulness they might have in the future.

**26.** Even if we did know what was going on inside the giant inscrutable matrices while the AGI was still too weak to kill us, this would just result in us dying with more dignity, if DeepMind refused to run that system and let Facebook AI Research destroy the world two years later. **Knowing that a medium-strength system of inscrutable matrices is planning to kill us, does not thereby let us build a high-strength system of inscrutable matrices that isn't planning to kill us.**

Yes to the bold part. It does tell us one machine *not* to build, it certainly *helps*, but it doesn't tell us how to fix the problem even if we get that test right somehow.

The non-bold part depends on the two-years thesis being true, but follows logically if you think that FAIR is always within two years of DeepMind and so on.

I cannot think of any death I want less than to be killed by Facebook AI research. Please, seriously, *anyone else*.

**27.** When you explicitly optimize against a detector of unaligned thoughts, you're partially optimizing for more aligned thoughts, and partially optimizing for unaligned thoughts that are harder to detect. **Optimizing against an interpreted thought optimizes against interpretability.**

Yes, obviously, I accidentally covered that already. I see why it had to be said out loud.

**28.** The AGI is smarter than us in whatever domain we're trying to operate it inside, so we cannot mentally check all the possibilities it examines, and we cannot see all the consequences of its outputs using our own mental talent. **A powerful AI searches parts of the option space we don't, and we can't foresee all its options.**

Yes to the bold text, obviously, and also yes to the implications by default.

If nothing else, an attempt to check the output of the AGI *means that we are checking the output of the AGI*, and as I noted previously that means it can communicate with humans, and it is a strong part of my core model that this should be assumed to be sufficient for a sufficiently generally powerful non-aligned AGI to manipulate the humans more generally, no matter the situation in any particular domain, although I can see bandwidth limitations that could make this less obvious slash raise the bar a lot for what would count as sufficiently powerful.

We can't check all the possibilities it examines, but is it obvious we can't see the consequences of its outputs using our own mental talent? That is potentially a fundamentally easier problem than generating or evaluating the possibilities.

Consider mathematics, a classic place people attempt to do something 'safe' with AGI. It is *much* easier to verify a proof than it is to generate that same proof, and requires a much lower level of intelligence and compute. It seems entirely plausible that the AGI is vastly better at math than Terrance Tao, can prove things in ways Tao didn't consider while occasionally cheating a bit on one of the steps, but Tao can still look over the proofs and say 'yes, that's right' when they are right and 'no, that's cheating' when they aren't, and be right.

There are plenty of more practical, more dangerous domains where that is also the case. Tons of problems are of the form 'There was essentially zero hope that I would have generated this course of action, but now that you propose it I understand what it would do and why it is or isn't a good idea.'

Nanotech and protein folding, which is used in the post as the canonical default unsafe thing to do, seem like areas where this is *not* the case. There are plenty of times when *by far* the most efficient thing to do, if you trust the AGI, is *not* to check all the consequences of its output, and it is highly plausible that pivotal acts require trusting the AGI in this way for all

solutions we have found so far. The existence of exceptions doesn't 'get us out' of the core problem here, but it seems important to be precise.

**29.** The outputs of an AGI go through a huge, not-fully-known-to-us domain (the real world) before they have their real consequences. **Human beings cannot inspect an AGI's output to determine whether the consequences will be good.**

Yes, obviously, for outputs that are sufficiently relevant to our interests here, and we can't use the ones where we *can* know the consequences to know what would happen when we can't. What we *can* potentially do with outputs is sometimes know *what those particular outputs* would do, at the cost of severe limitation, and also again we are reading outputs of an AGI which is a very bad idea if it isn't aligned.

**30. There is no pivotal output of an AGI that is humanly checkable and can be used to safely save the world but only after checking it;** this is another form of pivotal weak act which does not exist.

This is the rub of the whole section. There exist outputs that are humanly checkable. There exist outputs that are humanly checkable but not in practice humanly generatable. The claim is that no combination of such outputs can enable a pivotal act.

If true, then performing a pivotal act requires trusting the AGI, which means we will have to trust the AGI, despite having no reason to think this would be anything but the worst possible idea and no path to making it otherwise.

It is clear that no one has figured out how to avoid this, or at least no one willing to talk about it, despite quite a bit of trying. It is highly plausible that there is no solution. I continue not to be *convinced* there exists no solution.

I also know that if I thought I had such an act, it is highly plausible I would take one look at it and say 'I am *not* talking about that in public, absolutely not, no way in hell.'

**31.** A strategically aware intelligence can choose its visible outputs to have the consequence of deceiving you, including about such matters as whether the intelligence has acquired strategic awareness; **you can't rely on behavioral inspection to determine facts about an AI which that AI might want to deceive you about.** (Including how smart it is, or whether it's acquired strategic awareness.)

Yes, obviously. Same as a human, except (when it matters most) smarter about it. And anything internal you observe also becomes an output that it can do this on, as well.

**32.** Human thought partially exposes only a partially scrutable outer surface layer. Words only trace our real thoughts. Words are not an AGI-complete data representation in its native style. The underparts of human thought are not exposed for direct imitation learning and can't be put in any dataset. **This makes it hard and probably impossible to train a powerful system entirely on imitation of human words or other human-legible contents**, which are only impoverished subsystems of human thoughts; **unless that system is powerful enough to contain inner intelligences figuring out the humans**, and at that point it is no longer really working as imitative human thought.

Yes, except perhaps for the last bit after the bold.

Humans themselves contain inner intelligences figuring out humans. Relative to other tasks we are remarkably good at this one. If your goal was to train a powerful system, and your method was to have the system do so on language while in some sense figuring out the humans, that doesn't sound like it means you can't be imitating human thought? Especially since if the goal was to imitate human words, you'd potentially want to be *imitating the human interpretations of humans* rather than *correctly interpreting the humans*, as the

important thing, because you're trying to model what a human would have done next in text and that requires knowing what words would bubble out of their system rather than understanding what's *actually* going on around them.

**33. The AI does not think like you do,** the AI doesn't have thoughts built up from the same concepts you use, it is utterly alien on a staggering scale. Nobody knows what the hell GPT-3 is thinking, not *only* because the matrices are opaque, but because the *stuff within that opaque container* is, very likely, incredibly alien – nothing that would translate well into comprehensible human thinking, even if we could see past the giant wall of floating-point numbers to what lay behind.

Yes. The AI does not think like you do, and 99% of people need to understand this.

But maybe it *kind of* does? For two reasons.

One is that, again based on my discussions with Chris Olah, and another discussion I had with someone else working on interpretability, to the extent that they did look inside a giant inscrutable matrix it turned out to be surprisingly scrutable, and many of the neurons 'meant something.' That's not as helpful as one would hope, but it *is* an indication that some of the thinking isn't alien for the larger values of alien. It's still going to be *more alien* than any other humans are thinking, but the scale may not be so staggering in the end.

Which plays into the second reason, which is #22, the claim that there is a core function to general intelligence, which implies the possibility that in some sense we are Not So Different as all that. That's *compared to being completely alien and impossible to ever hope to decipher at all*, mind you, not compared to obvious nonsense like 'oh, you mean it's like how it's really hard to understand ancient Egyptians' or something, yes it is going to be a lot, lot more alien than that.

I continue to be skeptical that getting a general intelligence is that easy, but if it *is* that easy and follows this naturally, I wonder how much that implies it is (relatively) less alien.

## Section B.4

**34. Coordination schemes between superintelligences are not things that humans can participate in** (eg because humans can't reason reliably about the code of superintelligences); a "multipolar" system of 20 superintelligences with different utility functions, plus humanity, has a natural and obvious equilibrium which looks like "the 20 superintelligences cooperate with each other but not with humanity".

Yes. I am convinced that things like 'oh we will be fine because the AGIs will want to establish proper rule of law' or that we could somehow usefully be part of such deals are nonsense. I do think that the statement here on its own is *unconvincing* for someone not already convinced who isn't inclined to be convinced. I agree with it because I was already convinced, but unlike many points that should be shorter this one should have probably been longer.

**35. Schemes for playing "different" AIs off against each other stop working if those AIs advance to the point of being able to coordinate via reasoning about (probability distributions over) each others' code. Any system of sufficiently intelligent agents can probably behave as a single agent, even if you imagine you're playing them against each other.**

Yes. Not automatically or definitely, but enough of a probably that these plans are non-starters even if they weren't also non-starters for other reasons as well, which I think they are.

I at least somewhat remember when LessWrong was all about questions like this. That was a long time ago. A more civilized conversation from a more civilized age.

### **36. AI-boxing can only work on relatively weak AGIs; the human operators are not secure systems.**

Yes, the only way to keep a sufficiently strong AGI boxed is to not interact with its output in any way, and even then I wouldn't be so sure. Stop thinking there is hope here, everyone, please. Alas, my expectation is that the argument here is not going to be convincing to anyone who wasn't already convinced by now.

## **Section C**

Okay, those are some significant problems, but lots of progress is being made on solving them, right? There's a whole field calling itself "AI Safety" and many major organizations are expressing Very Grave Concern about how "safe" and "ethical" they are?

Yeah, it's not going so well. It is in fact going so incredibly poorly that so far the whole thing is quite plausibly vastly net negative, with most funding that has gone into "AI Safety" efforts serving as de facto capabilities research that both speeds things up and divides them and is only serving to get us killed faster. It is going so poorly that Eliezer is writing posts like this with *actual no idea* what useful things could be done, and when I ask people what could be done if one literally was directing policy for the President of the United States I got essentially no useful suggestions beyond trying to hire away GPU designers (or AI researchers if you have the budget for that) to design solar panels. Which, sure, better than not doing that but that is not a good answer.

**37.** There's a pattern that's played out quite often, over all the times the Earth has spun around the Sun, in which some bright-eyed young scientist, young engineer, young entrepreneur, proceeds in full bright-eyed optimism to challenge some problem that turns out to be really quite difficult. Very often the cynical old veterans of the field try to warn them about this, and the bright-eyed youngsters don't listen, because, like, who wants to hear about all that stuff, they want to go solve the problem! Then this person gets beaten about the head with a slipper by reality as they find out that their brilliant speculative theory is wrong, it's actually really hard to build the thing because it keeps breaking, and society isn't as eager to adopt their clever innovation as they might've hoped, in a process which eventually produces a new cynical old veteran. Which, if not literally optimal, is I suppose a nice life cycle to nod along to in a nature-show sort of way.

Sometimes you do something for the *first* time and there are no cynical old veterans to warn anyone and people can be *really* optimistic about how it will go; eg the initial Dartmouth Summer Research Project on Artificial Intelligence in 1956: "An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer."

This is *less* of a viable survival plan for your *planet* if the first major failure of the bright-eyed youngsters kills *literally everyone* before they can predictably get beaten about the head with the news that there were all sorts of unforeseen difficulties and reasons why things were hard. You don't get any cynical old veterans, in this case, because everybody on Earth is dead.

Once you start to suspect you're in that situation, you have to do the Bayesian thing and update now to the view you will predictably update to later: realize you're in a situation of being that bright-eyed person who is going to encounter Unexpected Difficulties later and end up a cynical old veteran – or would be, except for the part where you'll be dead

along with everyone else. And become that cynical old veteran *right away*, before reality whaps you upside the head in the form of everybody dying and you not getting to learn.

**Everyone else seems to feel that, so long as reality hasn't whapped them upside the head yet and smacked them down with the actual difficulties, they're free to go on living out the standard life-cycle and play out their role in the script and go on being bright-eyed youngsters; there's no cynical old veterans to warn them otherwise, after all, and there's no proof that everything won't go beautifully easy and fine, given their bright-eyed total ignorance of what those later difficulties could be.**

I mostly agree with the central thing that's being got at here in the end, but I think a lot of this is a misunderstanding of the proper role of Bright-Eyed Youngsters, so I want to kind of reason through this again.

If all the problems in the world were conveniently labeled with difficulty levels, or could be so assessed by the number of cynical old veterans sitting in their offices continuing to not solve the problem while writing enough papers to have tenure, and the way one solved problems was to accumulate Valuable Experience and Score Difficulty Points until the solving threshold was reached, then it would make sense that the purpose of a Bright-Eyed Youngster is to get smacked upside the head enough times to create a Cynical Old Veteran (COV). At which point perhaps they can make some progress and we can all praise the cycle of life.

Instead, I think the way that it works is that the COVs mostly *don't* solve such problems. Instead, the COVs are out of ideas of how to solve the problem, or have concluded the problem is hopeless, and write posts like Eliezer's about why the problem is doomed to never be solved. And they spend some of their time mentoring Bright-Eyed Youngsters, explaining to them why their ideas won't work and helping reality smack them upside the head more efficiently. When the youngster is actually on the right track, they often explain to them why their ideas are wrong anyway, and sometimes the youngster luckily does not listen. Also the veterans assign subproblems and determine who gets tenure.

Who actually solves problems? In general (not AGI specific) I am not going to bet on the Cynical Old Veterans too aggressively, especially the older and more cynical ones. Exactly how young or old to bet depends on the field - if AGI research is most similar to mathematics, presumably one should bet on quite young. If it's other things, less young, but I'd assume still rather young.

You should update straight to 'this particular problem of building an AGI is super difficult' without requiring failed attempts, through reasoning out the nature of the problem, but my hunch is you want to in some senses remain a BEY anyway.

The bright-eyed thing is a feature (and the young thing is *definitely* a feature), because they make people actually try to solve problems for real. Most people don't react to learning that AGI is as hard as it is (if they do ever learn that) by saying 'all right, time to score as many dignity points as possible and work on the actually hard parts of this problem' instead they either find a way to unlearn the thing as quietly and quickly as possible, or they ignore it and keep publishing, or they go do something else, or they despair. That's typical, if you tell me a problem is impossible chances are I'll find something else to do or start doing fake work. A response of 'yes this is an impossible problem but I'll solve it anyway' seems great.

The structure implies any given unsolved problem is hard, including *for new problems*. Which doesn't seem right in general - this particular problem is indeed hard but many unsolved problems seem hard to COVs but are easy in the face of an actual attempt. Often when you start on a new problem it turns out it really is easy, because there's no selection against it being easy. Many problems turn out to be *shockingly* easy in the face of a real attempt. It is exactly the youngsters who think the problem is easy because they see something unique about it that are most likely to actually solve it, even though they're still presumably not realizing how hard it is, the same way that start-up founders usually have no idea what

they're signing up for but also that's how they actually found start-ups. Which, when they work, then proceed to use reality to slap the COVs upside the head on the way out. Or science can advance one funeral at a time.

The difference here is that a Bright-Eyed Youngster (BEY) working on most problems will waste some resources but doesn't do much real harm. In AGI there's the danger they will literally kill everyone on the planet. That's new.

So far they haven't killed everyone, but also BEYs are also failing to turn into skilled COVs because they don't even have the opportunity to properly fail (and kill everyone).

This does require some adjustments, especially once a BEY could potentially build an AGI. There's some confusion here if the BEY is thinking they know how to do *safety* versus thinking they know how to do an AGI at all (the most BEY of the BEYs don't even realize safety is a problem) but mostly this still should refer to safety.. At which point, yes, you very much don't want to trust that BEY's safety idea, and if they want to succeed at safety they need to be able to do it without being told by reality that their first few answers were hopelessly naïve.

This could be an argument that you want to use more veteran people, who have a relatively bigger sense of these issues. They have a better relative chance to actually solve the problem in this situation. Failure to previously solve it isn't evidence against them, because the problem won't up until then have been something that could potentially be solved, and error correction is relatively important. When I became a Cynical Old Veteran of Magic: The Gathering, I was *much* better about getting things right *on the first try* than I used to be, while simultaneously being worse at truly innovating. Which may or may not be the trade-off you need.

The report is that true worthwhile COVs (other than Eliezer) don't exist, there's no one else sitting around *not* pretending to do fake things but happy to teach you exactly why you'll fail. Or so the report goes..

The Bayesian point stands. Ideally a BEY should update on a problem not having been solved despite much effort and conclude it is likely very hard, and not hide from all the particular things that need to be dealt with, yet continue to have the enthusiasm to work on the problem while behaving in useful ways *as if the problem will turn out to be easy for them in particular* for some reason, if by 'easy' we mean just barely solvable, without actually believing that they will solve it.

Everyone being killed on the first attempt to solve the problem doesn't tell you the difficulty level of the problem *aside from the fact that the first failed attempt kills everyone*. This seems like it goes double if in order to try and solve the problem you first need to solve another problem that is just now becoming solvable, since you can't have a safe AGI without a way to make an AGI to begin with. So you have to think about the problem and figure it out that way.

So yes, young warrior, you must forge a Sword of Good Enough and take it into the Dungeon of Ultimate Evil and find your way to the evil wizard and slay him. But if you take an actual Sword of Good Enough in and the wizard gets it, that's it, everyone dies, world over. It's *probably* going to involve overwhelming odds against you, I mean did you see the sign above the dungeon or hear the screams inside, things look pretty grim, but our evidence is based on reasoning out what is logically going to be in this high level a dungeon, because we've never had anyone run into the dungeon with an actual Sword of Good Enough and get smacked upside the head by reality, and we know this because if they had we'd all be dead now.

And you can't wait forever, because there are plenty of other people who think they're heroes in a video game with save points and are going to try and speed run the damn thing,

and it won't be that long before one of them figures out how to forge a sword and gets us all killed, so 'grind an absurd amount before entering' means you never get a chance at all.

If there were a bunch of dead heroes to point to and people who ran away screaming to save their lives, then you could say 'oh I guess I should update that this dungeon is pretty tough' but without them the others get to fool themselves into thinking it might be that easy, and if it is then getting there late won't get them the glory.

I remember starting my own start-up as a BEY (except founder, not researcher), noticing the skulls, and thinking the problem was almost certainly incredibly hard and also probably much harder than I thought it was (but much less more hard than my estimates than the gap for most founders, and I think this proved true although our *particular* idea was bad and therefore unusually hard), and also that so what I had odds let's do this anyway, and then I went out and did it again as more of a hybrid with a better idea that was relatively easier, but same principle. That doesn't apply here, because there were attempts that went anywhere at all even at fully unsafe AGIs, and thus no failures or successes, resulting in zero successes but also zero veterans and zero skulls.

The problem comes from the BEY getting us all killed, by *actually attempting to win the game* via a half-baked solution that has zero chance of working on multiple levels, in a way that would normally not matter but here is deadly because an AGI is involved. And sure, point taken, but as long as that's not involved what's the problem with BEYs going in and boldly working on new safety models only to have reality smack them upside the face a lot?

My Eliezer model says that what's wrong with that is that *this causes them to do fake research*, in the sense that it isn't *actually* trying to solve the problem slash has zero chance of being helpful except insofar as it has a chance of teaching them enough to turn them into cynical veterans, and there isn't enough feedback to make them into veterans because reality isn't going to smack them upside the head strongly enough until it actually kills everyone.

And also the problem that *most things people tell themselves are safety work are actually capability work* and thus if you are not *actually* doing the hard safety work you are far more likely to advance capability and make things worse than you are to have some amazing breakthrough.

Or even worse, the problem is that the BEYs will *actually succeed at the fake problem* of alignment that *looks* like it would work that they actually think they've solved it and they are willing to turn on an AGI.

Thus, what you actually need is a BEY who is aware of why the problem is impossible (in the shut up and do the impossible sense) and thus starts work on the *real* problems, and everyone else is far worse than worthless because of what we know about the shape of the problem and how people interact with it and what feedback it gives us – assuming that our beliefs on this are correct, and I say 'our' because I mostly think Eliezer is right.

Notice the implications here. If the premises here are correct, and I believe they probably are, they seem to imply that 'growing the field' of AI Safety, or general 'raising awareness' of AI Safety, is quite likely to be *an actively bad idea*, unless they lead to things that will help, which means either (A) people who actually get what they're facing and/or (B) people who try to stop or slow down AGI development rather than trying to make it safer.

**38. It does not appear to me that the field of 'AI safety' is currently being remotely productive on tackling its enormous lethal problems.** These problems are in fact out of reach; the contemporary field of AI safety has been selected to contain people who go to work in that field anyways. Almost all of them are there to tackle problems on which they can appear to succeed and publish a paper claiming success; if they can do that and get funded, why would they embark on a much more unpleasant project of trying something harder that they'll fail at, just so the human species can die

with marginally more dignity? This field is not making real progress and does not have a recognition function to distinguish real progress if it took place. You could pump a billion dollars into it and it would produce mostly noise to drown out what little progress was being made elsewhere.

Yes, and again, *it seems like this is not saying the quiet part out loud*. The quiet part is 'I say not being productive on tackling lethal problems but what I actually meant is they are making our lethal problems worse by accelerating them along and letting people fool themselves about the lethality of those problems, so until we have a better idea please stop.'

**39. I figured this stuff out using the [null string](#) as input**, and frankly, I have a hard time myself feeling hopeful about getting real alignment work out of somebody who previously sat around waiting for somebody else to input a persuasive argument into them. This ability to "notice lethal difficulties without Eliezer Yudkowsky arguing you into noticing them" currently is an opaque piece of cognitive machinery to me, I do not know how to train it into others. It probably relates to '[security mindset](#)', and a mental motion where you refuse to play out scripts, and being able to operate in a field that's in a state of chaos.

Security mindset seems highly related, and the training thing here seems like it shouldn't be that hard? Certainly it seems very easy compared to the problem the trained people will then need to solve, and I think Eliezer has de facto trained me a substantial amount in this skill through examples over the years. There was a time I didn't have security mindset at all, and now I have at least *some* such mindset, and *some* ability to recognize lethal issues others are missing. He doesn't say how many other people he knows who have the abilities referred to here, I'd be curious about that. Or whether he knows anyone who has acquired them over time.

If the class 'AI researcher without this mindset' is net negative, and one with it is net positive, then we need to get CFAR and/or others on the case. This problem seems more like 'not that many people have made a serious attempt and it seems quite likely to be not impossible' than 'this seems impossible.'

If nothing else, a substantial number of other people do have security mindset, and you can presumably find them by looking at people who work in security, and presumably a bunch of them have thought about how to teach it?

**40. "Geniuses"** with nice legible accomplishments in fields with tight feedback loops where it's easy to determine which results are good or bad right away, and so validate that this person is a genius, are (a) people who might not be able to do equally great work away from tight feedback loops, (b) people who chose a field where their genius would be nicely legible even if that maybe wasn't the place where humanity most needed a genius, and (c) probably don't have the mysterious gears simply because they're *rare*.

**You cannot just pay \$5 million apiece to a bunch of legible geniuses from other fields and expect to get great alignment work out of them.**

They probably do not know where the real difficulties are, they probably do not understand what needs to be done, *they cannot tell the difference between good and bad work*, and the funders also can't tell without me standing over their shoulders evaluating everything, which I do not have the physical stamina to do.

I concede that real high-powered talents, especially if they're still in their 20s, genuinely interested, and have done their reading, are people who, yeah, fine, have higher probabilities of making core contributions than a random bloke off the street. But I'd have more hope – not significant hope, but *more* hope – in separating the concerns of (a) credibly promising to pay big money retrospectively for good work to anyone who

produces it, and (b) venturing prospective payments to somebody who is predicted to maybe produce good work later.

The problem with promising to pay big money retrospectively for good work is that, while an excellent idea, it doesn't actually solve the motivation problem if the problem with getting 'good work' out of people is that the probability of success for 'good work' is very low.

Which is indeed the problem, as I understand Eliezer describing it and I think he's largely right. Someone who enters the field who chooses to do real work has to recognize the need for 'real' work (he calls it 'good' above, sure), know what real work is and how to do it, and choose to attempt real work despite knowing that the default outcome that probably happens is that no good work results and thus the payoff is zero.

That is, unless there is some way to recognize a real failed *attempt* to do real work and reward *that*, but we don't have a hopeful path for accurately doing that without actual Eliezer doing it, for which the stamina is unavailable..

The question then is, sure, paying the \$5 million isn't super likely to get good work out of any individual person. But it's at least *kind of* true that we have billions of dollars that wants to be put to work on AI Safety, that isn't being spent because it can't help but notice that spending more money on current AI Safety options isn't going to generate positive amounts of dignity, and in fact likely generates negative amounts.

The real potential advantage of the \$5-million-to-the-genius approach is not that the genius is a favorite to do useful work. The advantage is that if you select such people based on them understanding the true difficulty of the problem, which is reinforced by the willingness to cut them the very large check and also the individual attention paid to them before and after check writing to ensure they 'get it,' they may be likely to *first, do no harm*. It seems plausible, at least, that they would 'fail with dignity' when they inevitably fail, in ways that don't make the situation *worse*, because they are smart enough to at least not do that.

So you could be in a situation where paying 25 people \$200k ends up being worse than doing nothing, while paying one promising genius \$5 million is at least *better* than doing nothing. And given the value of money versus the value of safety work, it's a reasonable approximation to say that anything with positive value is worth spending a lot of money. If the bandwidth required has rival uses that's another cost, but right now the alternative uses might be things we are happy to stop.

Another theory, of course, is that *introducing a genius to the questions surrounding AGI* is a *deeply, deeply foolish* thing to be doing. Their genius *won't obviously transfer* to knowing not to end up doing capabilities work or accidentally having (and sharing) good capabilities ideas, so the last thing you want to do is take the most capable people in the world at figuring things out and have them figure out the thing you least want anyone to figure out.

As far as I can tell, that's the real crux here, and I don't know which side of it is right?

#### **41. Reading this document cannot make somebody a core alignment researcher.**

**researcher.** That requires, not the ability to read this document and nod along with it, but the ability to spontaneously write it from scratch without anybody else prompting you; that is what makes somebody a peer of its author. It's guaranteed that some of my analysis is mistaken, though not necessarily in a hopeful direction.

The ability to do new basic work noticing and fixing those flaws is the same ability as the ability to write this document before I published it, which nobody apparently did, despite my having had other things to do than write this up for the last five years or so.

Some of that silence may, possibly, optimistically, be due to nobody else in this field having the ability to write things comprehensibly – such that somebody out there had the knowledge to write all of this themselves, if they could only have written it up, but

they couldn't write, so didn't try. I'm not particularly hopeful of this turning out to be true in real life, but I suppose it's one possible place for a "positive model violation" (miracle). The fact that, twenty-one years into my entering this death game, seven years into other EAs noticing the death game, and two years into even normies starting to notice the death game, it is still Eliezer Yudkowsky writing up this list, says that humanity still has only one gamepiece that can do that. I knew I did not actually have the physical stamina to be a star researcher, I tried really really hard to replace myself before my health deteriorated further, and yet here I am writing this.

That's not what surviving worlds look like.

Yes, mostly. A lot of distinct claims to unpack here, which is why it is quoted in full.

Reading this document is different from being able to understand and recreate the arguments, or the ability to generate additional similar arguments on things that weren't mentioned or in response to new objections or ideas.

The bolder claim is the idea that if you couldn't have written something similar to this document yourself, you can't usefully research AI Safety.

(Notice once again that this is saying that almost no one can usefully research AI Safety and that we'd likely be better off if most of the people doing so stopped trying, or at least/most worked on first becoming able to generate such a document rather than directly on the problem.)

On the question of writing ability?

I will say outright that yes, that is an important barrier here.

The chance of any given person, who could have otherwise generated the list, lacking the required writing ability. Writing ability on the level of Eliezer isn't as rare as understanding of the problem on the level of Eliezer, but it is quite rare. How many people would have a similar chance to Eliezer of 'pulling off' HPMOR or the sequences purely in terms of writing quality, even if they understood the core material about as well?

Writing the list *in this way* is a thing Eliezer gets to do that others mostly don't get to do. If *someone else* wrote up the list with this level of ranting and contempt, I would not expect that to go well, and that would reasonably lead someone else capable of writing it that way to not do so.

The job of someone else writing this list *properly* is much harder. They would feel the need to write it 'better' in some ways which would make it longer, and also probably make it worse for at least several iterations. The job of *deciding to write it* is much harder, requiring the author to get past a bunch of social barriers and modesty issues and so on. At best it would not be a fast undertaking.

One could reasonably argue that there's a strong *anti*-correlation in skills here. How do you get good at writing? You write. A lot. All the time. There are no substitutions. And that's a big time commitment.

So how many people in the broad AI Safety have *written enough words* in the right forms to plausibly have the required writing ability here even in theory? There are at most a handful.

And of course, writing such a list is not a normal default social action so it doesn't happen, and even Eliezer took forever to actually write the list and post it, and ended up deciding to post a self-described subpar version for want of ability to write a good one, despite knowing how important such a thing was and having all the required knowledge.

That does not mean there are people who, if imbued with the writing skill, could have written the list. It simply means we don't have the Bayesian evidence to know.

I agree that, in the cases where Eliezer is right about the nearness and directness of the path to AGI, this is mostly not what surviving worlds look like, but also I've learned that everyone everywhere is basically incompetent at everything and also not trying to do it in the first place, and yet here we still are, so let's not despair too much every time we get that prior confirmed again. If you told me a lot of the things I know now ten years ago I'd have also said 'that's not what surviving civilizations look like' purely in terms of *ordinary* ruin.

**42. There's no plan.** Surviving worlds, by this point, and in fact several decades earlier, have a plan for how to survive. It is a written plan. The plan is not secret. In this non-surviving world, there are no candidate plans that do not immediately fall to Eliezer instantly pointing at the giant visible gaping holes in that plan. Or if you don't know who Eliezer is, you don't even realize you need a plan, because, like, how would a human being possibly realize that without Eliezer yelling at them?

Yes, there is no plan. I would like to have a plan. Not having *any plan at all*, of *any* detail, that offers a path forward, is indeed not what surviving worlds usually look like.

Yet I am not convinced that surviving worlds involve a plan along the lines above.

You know who else doesn't have a plan that Eliezer (you'd think I would say whoever the domain-equivalent of Eliezer is and that would work too, but honestly literal Eliezer would mostly work fine anyway) couldn't point at the visible gaping holes in?

Yeah, with notably rare exceptions the answer is actual everyone else.

I do realize that the whole point is that the kind of complete incompetence and muddling through via trial and error we usually do won't work on this one, so that offers little comfort in some sense, but the visible written plan that actually works available decades in advance is not how humans work. If anything, this feels like one of those reality-falsifying assumptions Eliezer is (wisely) warning everyone else *not* to make about other aspects of the problem, in the sense that this is trying to make the solution run through a plan like that which *kind of* is like assuming such a plan could possibly exist. Which in turn seems like it is either a very bold claim about the nature of humanity and planning, the nature of the problem and solution space (in a way that goes in a very different direction than the rest of the list), or more likely both.

This document wasn't written until well after it *could* have been written by Eliezer. Part of that is health issues, but also part of that clearly is that we wasted a bunch of time thinking we'd be able to offer better ideas and better plans and thus didn't proceed with worse slash less ready ideas and plans as best we could. The new plan of not holding out as much for a better plan is indeed a better, if highly non-ideal, plan.

Relatively few are aware even that they should, to look better, produce a *pretend* plan that can fool EAs too '[modest](#)' to trust their own judgments about seemingly gaping holes in what serious-looking people apparently believe.

Is this right? Should I have produced a pretend plan? Should I be pretending to write one here and now? Actually writing a bad one? How many people should have produced one? Do we want to look better?

If everyone is being overly modest (and mostly they are) then there's also a big danger of information cascades during this kind of creation of common knowledge. Everyone converging around our failure to make any progress and the situation being grim seems clearly right to me. Everyone converging around many other aspects of the problem space worries me more as I am not convinced by the arguments.

**43. This situation you see when you look around you is not what a surviving world looks like.** The worlds of humanity that survive have plans. They are not leaving to one tired guy with health problems the entire responsibility of pointing out real and lethal problems proactively. Key people are taking internal and real responsibility for finding flaws in their own plans, instead of considering it their job to propose solutions and somebody else's job to prove those solutions wrong. That world started trying to solve their important lethal problems earlier than this. Half the people going into string theory shifted into AI alignment instead and made real progress there. When people suggest a planetarily-lethal problem that might materialize later - there's a lot of people suggesting those, in the worlds destined to live, and they don't have a special status in the field, it's just what normal geniuses there do - they're met with either solution plans or a reason why that shouldn't happen, not an uncomfortable shrug and 'How can you be sure that will happen' / 'There's no way you could be sure of that now, we'll have to wait on experimental evidence.'

A lot of those better worlds will die anyways. It's a genuinely difficult problem, to solve something like that on your first try. But they'll die with more dignity than this.

I go back and forth on what my relationship should be to the problem of AI Safety, and what the plan should be to address it both on a personal and general strategic level. I've come around largely to the perspective that my comparative advantage mostly lies elsewhere, and that many other aspects of our situation are both threatening to doom us even without or before AGI dooms us and also even their lesser consequences are why our world looks like it does (as in: not like one that is that likely to survive AGI when it happens). So it makes sense for me to mostly work on making the world/civilization more generally look like one that gets to survive in many ways, rather than directly attack the problem.

At other times I get to wondering if maybe I *should* try to tackle the problem directly based on having been able to usefully attempt tackling of problems I should have had no business attempting to tackle. I do reasonably often get the sense that these problems have solutions and with the right partnerships and resources I could be able to have a chance of finding them. Who knows.

## Conclusion

I put the core conclusions at the *top* rather than the bottom, on the theory that many/most people quite reasonably won't read this far. I was on the fence, before being funded to do it, on whether writing this was a good idea given my current level of domain knowledge and the risk of wasting not only my own but other people's time. Having written it, it seems like it is plausibly useful, so hopefully that turns out to be right. There are various secondary documents that could be produced that require a combination of writing skill and understanding of the problem and willingness to go ahead and write drafts of them, and it is not crazy that I might be the least terrible solution to that problem for some of them.

# Why all the fuss about recursive self-improvement?

*This article was outlined by Nate Soares, inflated by Rob Bensinger, and then edited by Nate. Content warning: the tone of this post feels defensive to me. I don't generally enjoy writing in "defensive" mode, but I've had this argument thrice recently in surprising places, and so it seemed worth writing my thoughts up anyway.*

---

In last year's [Ngo/Yudkowsky conversation](#), one of Richard's big criticisms of Eliezer was, roughly, 'Why the heck have you spent so much time focusing on recursive self-improvement? Is that not indicative of poor reasoning about AGI?'

I've heard similar criticisms of MIRI and FHI's past focus on orthogonality and instrumental convergence: these notions seem obvious, so either MIRI and FHI must be totally confused about what the big central debates in AI alignment are, or they must have some very weird set of beliefs on which these notions are somehow super-relevant.

This seems to be a pretty common criticism of past-MIRI (and, similarly, of past-FHI); in the past month or so, I've heard it two other times while talking to other OpenAI and Open Phil people.

This argument looks misguided to me, and I hypothesize that a bunch of the misguidedness is coming from a simple failure to understand the relevant history.

I joined this field in 2013-2014, which is far from "early", but is early enough that I can attest that recursive self-improvement, orthogonality, etc. were geared towards a *different argumentative environment*, one dominated by claims like "AGI is impossible", "AGI won't be able to exceed humans by much", and "AGI will naturally be good".

A possible response: "Okay, but 'sufficiently smart AGI will recursively self-improve' and 'AI isn't automatically nice' are still *obvious*. You should have just ignored the people who couldn't immediately see this, and focused on the arguments that would be relevant to hypothetical savvy people in the future, once the latter joined in the discussion."

I have some sympathy for this argument. Some considerations weighing against, though, are:

- I think it makes more sense to filter on argument validity, rather than "obviousness". What's obvious varies a lot from individual to individual. If just about everyone talking about AGI is saying "obviously false" things (as was indeed the case in 2010), then it makes sense to at least *try* publicly writing up the obvious counter-arguments.
- This seems to assume that the old arguments (e.g., in *Superintelligence*) didn't *work*. In contrast, I think it's quite plausible that "everyone with a drop of sense in them agrees with those arguments today" is true in large part *because* these propositions were explicitly laid out and argued for in the past. The claims we take as background now are the claims that were fought for by the old guard.

- I think this argument overstates how many people in ML today grok the “obvious” points. E.g., based on a recent [DeepMind Podcast episode](#), these sound like likely points of disagreement with [David Silver](#).

But even if you think this was a strategic error, I still think it’s important to recognize that MIRI and FHI were *arguing correctly against the mistaken views of the time*, rather than *arguing poorly against future views*.

## Recursive self-improvement

Why did past-MIRI talk so much about recursive self-improvement? Was it because Eliezer was super confident that humanity was going to get to AGI via the route of a seed AI that understands its own source code?

I doubt it. My read is that Eliezer did have “seed AI” as a top guess, back before the deep learning revolution. But I don’t think that’s the main source of all the discussion of recursive self-improvement in the period around 2008.

Rather, my read of the history is that MIRI was operating in an argumentative environment where:

- Ray Kurzweil was claiming things [along the lines of](#) ‘Moore’s Law will continue into the indefinite future, even past the point where AGI can contribute to AGI research.’ (The [Five Theses](#), in 2013, is a list of the key things Kurzweilians were getting wrong.)
- Robin Hanson was claiming things [along the lines of](#) ‘The power is in the culture; superintelligences wouldn’t be able to outstrip the rest of humanity.’

The memetic environment was one where most people were either ignoring the topic altogether, or asserting ‘AGI cannot fly all that high’, or asserting ‘AGI flying high would be business-as-usual (e.g., with respect to growth rates)’.

The weighty conclusion of the “recursive self-improvement” meme is not “expect seed AI”. The weighty conclusion is “sufficiently smart AI will rapidly improve to heights that leave humans in the dust”.

Note that this conclusion is still, to the best of my knowledge, *completely true*, and recursive self-improvement is a *correct argument for it*.

Which is not to say that recursive self-improvement happens before the end of the world; if the first AGI’s mind is sufficiently complex and kludgy, it’s entirely possible that the cognitions it implements are able to (e.g.) crack nanotech well enough to kill all humans, before they’re able to crack themselves.

The big update over the last decade has been that humans might be able to fumble their way to AGI that can do crazy stuff *before* it does much self-improvement.

(Though, to be clear, from my perspective it’s still entirely plausible that you will be able to turn the first general reasoners to their own architecture and get a big boost, and so there’s still a decent chance that self-improvement plays an important early role. (Probably destroying the world in the process, of course. Doubly so given that I expect it’s even harder to understand and align a system if it’s self-improving.))

In other words, it doesn't seem to me like developments like deep learning have undermined the recursive self-improvement argument in any real way. The argument seems solid to me, and reality seems quite consistent with it.

Taking into account its past context, recursive self-improvement was a *super conservative* argument that has been *vindicated in its conservatism*.

It was an argument for the proposition "AGI will be able to exceed the heck out of humans". And AlphaZero came along and was like, "Yep, that's true."

Recursive self-improvement was a super conservative argument for "AI blows past human culture eventually"; when reality then comes along and says "yes, this happens in 2016 when the systems are far from truly general", the update to make is that this way of thinking about AGI sharply outperformed, not that this way of thinking was silly because it talked about sci-fi stuff like recursive self-improvement when it turns out you can do crazy stuff without even going that far. As Eliezer [put it](#), "reality held a more extreme position than I did on the Yudkowsky-Hanson spectrum".

If arguments like recursive self-improvement and orthogonality seem irrelevant and obvious now, then great! Intellectual progress has been made. If we're lucky and get to the next stop on the train, then I'll hopefully be able to link back to this post when people look back and ask why we were arguing about all these other silly obvious things back in 2022.

## Deep learning

I think "MIRI staff spent a bunch of time talking about instrumental convergence, orthogonality, recursive self-improvement, etc." is a silly criticism.

On the other hand, I think "MIRI staff were slow to update about how far deep learning might go" is a fair criticism, and we lose Bayes points here, especially relative to people who were vocally bullish about deep learning before late 2015 / early 2016.

In 2003, deep learning didn't work, and nothing else worked all that well either. A reasonable guess was that we'd need to understand intelligence in order to get unstuck; and if you understand intelligence, then an obvious way to achieve superintelligence is to build a simple, small, clean AI that can take over the hard work of improving itself. This is the idea of "seed AI", as I understand it. I don't think 2003-Eliezer thought this direction was certain, but I think he had a bunch of probability mass on it.<sup>[1]</sup>

I think that Eliezer's model was somewhat surprised by humanity's subsequent failure to gain much understanding of intelligence, and also by the fact that humanity was able to find relatively brute-force-ish methods that were computationally tractable enough to produce a lot of intelligence anyway.

But I also think this was a reasonable take in 2003. Other people had even better takes — Shane Legg comes to mind. He stuck his neck out early with narrow [predictions](#) that panned out. Props to Shane.

I personally had run-of-the-mill bad ideas about AI as late as 2010, and didn't turn my attention to this field until about 2013, which means that I lost a bunch of Bayes

points relative to the people who managed to figure out in 1990 or 2000 that AGI will be our final invention. (Yes, even if the people who called it in 2000 were expecting seed AI rather than deep learning, back when nothing was really working. I reject the [Copenhagen Theory](#) Of Forecasting, according to which you gain special epistemic advantage from not having noticed the problem early enough to guess wrongly.)

My sense is that MIRI started taking the deep learning revolution much more seriously in 2013, while having reservations about whether broadly deep-learning-like techniques would be the first way humanity reached AGI. Even now, it's not completely obvious to me that this will be the broad paradigm in which AGI is first developed, though something like that seems fairly likely at this point. But, if memory serves, during the Jan. 2015 Puerto Rico conference I was treating the chance of deep learning going all the way as being in the 10-40% range; so I don't think it would be fair to characterize me as being totally blindsided.

My impression is that Eliezer and I, at least, updated harder in 2015/16, in the wake of AlphaGo, than a bunch of other locals (and I, at least, think I've been less surprised than various other vocal locals by GPT, PaLM, etc. in recent years).

Could we have done better? Yes. Did we lose Bayes points? Yes, especially relative to folks like Shane Legg.

But since 2016, it mostly looks to me like with each AGI advancement, others update towards my current position. So I'm feeling pretty good about the predictive power of my current models.

Maybe this all sounds like revisionism to you, and your impression of FOOM-debate-era Eliezer was that he loved GOFAI and thought recursive self-improvement was the only advantage digital intelligence could have over human intelligence.

And, I wasn't here in that era. But I note that Eliezer [said the opposite](#) at the [time](#); and the track record for such claims seems to hold more examples of "mistakenly rounding the other side's views off to a simpler, more-cognitively-available caricature", and fewer examples of "peering past the veil of the author's text to see his hidden soul".

Also: It's important to ask proponents of a theory what they predict will happen, before crowing about how their theory made a misprediction. You're always welcome to ask for my predictions in advance.

(I've been making this offer to people who disagree with me about whether I have egg on my face since 2015, and have rarely been taken up on it. E.g.: yes, we too predict that it's easy to get GPT-3 to tell you the answers that humans label "aligned" to simple word problems about what we think of as "ethical", or whatever. That's never where we thought the difficulty of the alignment problem was in the first place. Before saying that this shows that alignment is actually easy contra everything MIRI folk said, consider asking some MIRI folk for their predictions about what you'll see.)

## 1. ^

In particular, I think Eliezer's best guess was AI systems that would look small, clean, and well-understood relative to the large opaque artifacts produced by deep learning. That doesn't mean that he was picturing GOFAI; there exist a

wide range of possibilities of the form “you understand intelligence well enough to not have to hand off the entire task to a gradient-descent-ish process to do it for you” that do not reduce to “coding everything by hand”, and certainly don’t reduce to “reasoning deductively rather than probabilistically”.

# LessWrong Has Agree/Disagree Voting On All New Comment Threads

Starting today we're activating two-factor voting on all new comment threads.

Now there are two axes on which you can vote on comments: the standard karma axis remains on the left, and the new axis on the right lets you show much you agree or disagree with the content of a comment.

Mod note: I activated two-axis voting on this post, since it seemed like it would make the conversation go better.  
Reply

I agree.  
Reply

## How the system works

For the pre-existing voting system, the most common interpretation of up/down-voting is "Do I want to see more or less of this content on the site?" As an item gets more/less votes, the item changes in visibility, and the karma-weighting of the author is eventually changed as well.

Agree/disagree is just added on to this system. Here's how it all hooks up.

- **Agree/disagree voting does not translate into a user's or post's karma — its sole function is to communicate agreement/disagreement.** It has no other direct effects on the site or content visibility (i.e. no effect on sorting algorithms).
- For both regular voting and the new agree/disagree voting, you have the ability to normal-strength vote and strong-vote. Click once for normal-strength vote. For strong-vote, click-and-hold on desktop or double-tap on mobile. The weight of your strong-vote is approximately proportional to your karma on a log-scale ([exact numbers here](#)).

## Ben's personal reasons for being excited about this split

Here's a couple of reasons that are alive for me.

- I personally feel much more comfortable upvoting good comments that I disagree with or whose truth value I am highly uncertain about, because I don't feel that my vote will be mistaken as setting the social reality of what is true.
- I also feel very comfortable strong-agreeing with things while not up/downvoting on them, so as to indicate which side of an argument seems true to me without my voting being read as "this person gets to keep accruing more and more social status for just repeating a common position at length".
- Similarly to the first bullet, I think that many writers have *interesting* and *valuable* ideas but whose truth-value I am quite unsure about or even disagree with. This split allows voters to repeatedly signal that a given writer's comments are of high value, without building a false-consensus that LessWrong has high confidence that the ideas

are true. (For example, many people have incompatible but valuable ideas about how AGI development will go, and I want authors to get lots of karma and visibility for excellent contributions without this ambiguity.)

- There are many comments I think are bad but am averse to downvoting, because I feel that it is ambiguous whether the person is being downvoted because everyone thinks their take is unfashionable or whether it's because the person is wasting the commons with their behavior (e.g. belittling, starting bravery debates, not doing basic reading comprehension, etc). With this split I feel more comfortable downvoting bad comments without worrying that everyone else who states the position will worry if they'll also be downvoted.
- I have seen some comments that previously would have been "downvoted to hell" are now on positive karma, and are instead "disagreed to hell". I won't point them out to avoid focusing on individuals, but this seems like an obvious improvement in communication ability.

I could go on but I'll stop here.

## Please give us feedback

This is one of the main voting experiments we've tried on the site ([here's the other one](#)). We may try more changes and improvement in the future. Please let us know about your experience with this new voting axis, especially in the next 1-2 weeks.

If you find it concerning/invigorating/confusing/clarifying/other, we'd like to know about it. Comment on this post with feedback and I'll give you an upvote (and maybe others will give you an agree-vote!) or let us know in the intercom button in the bottom right of the screen.

We've rolled it out on many (15+) threads now ([example](#)), and my impression is that it's worked as hoped and allowed for better communication about the truth.

The screenshot shows a comment by Matthew Barnett. At the top, there are navigation buttons: a back arrow, the author's name 'Matthew Barnett', a timestamp '8d', a profile icon, a reply count '28', a left arrow, a right arrow, and a close button. Below the buttons, the comment text reads: 'It's as good as time as any to re-iterate my reasons for disagreeing with what I see as the Yudkowskian view of future AI. What follows isn't intended as a rebuttal of any specific argument in this essay, but merely a pointer that I'm providing for readers, that may help explain why some people might disagree with the conclusion and reasoning contained within.' Underneath the text, it says 'I'll provide my cruxes point-by-point,' followed by a bulleted list: '• I think raw intelligence, while important, is not the primary factor that explains why humanity-as-a-species is much more powerful than chimpanzees-as-a-species. Notably, humans were once much less powerful, in our hunter-gatherer days, but over time, through the [gradual process of accumulating technology, knowledge, and culture](#), humans now possess vast productive capacities that far outstrip our ancient powers.' Below this, another paragraph begins: 'Similarly, our ability to coordinate through language also plays a huge role in explaining our power compared to other animals. But, on a first approximation, other animals can't coordinate *at all*, making this distinction much less impressive. The first AGIs we construct will be born into a culture already capable of coordinating, and sharing knowledge, making the potential power di' At the bottom of the comment, there is a link '... (read more)' and a 'Reply' button.

One virtue of the new voting axis is how it allows users to clearly express great appreciation for high-quality comments that many users disagree with.

# The prototypical catastrophic AI action is getting root access to its datacenter

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

(I think Carl Shulman came up with the “hacking the SSH server” example, thanks to him for that. Thanks to Ryan Greenblatt, Jenny Nitishinskaya, and Ajeya Cotra for comments.)

EDIT: I recommend reading my discussion with Oli in the comments for various useful clarifications.

In my opinion, the prototypical example of an action which an AI can take which is unacceptably bad even if it happens very rarely (aka a catastrophic failure) is something like “the AI makes a code change to some code used on the datacenter where it is running that adds a backdoor, allowing it to get root access to all the computers in the datacenter”. As a full story, this looks something like:

At some point when the AI is quite powerful and useful but not totally intellectually self-sufficient yet, the AI makes a code change to some code used on the datacenter where it is running that adds a backdoor, allowing it to get root access to all the computers in the datacenter. Once it has root access, it modifies the SSH server so that whenever any humans log into the datacenter, the AI is able to intercept any of their commands or the data that they see. The AI gets increasingly powerful over the next month, as the human ML researchers with AI assistance continue to make progress (unaware that some things that they’re seeing are faked) and various training runs/self-improvement loops continue. After a month the AI is powerful enough that it decides it’s worth it to take overtly aggressive actions to ensure the physical security of its data centers and allow it to build lots of factories; this is probably the point at which we die, if we die.

This is in contrast to a catastrophic action like:

The AI figures out how to develop Drexlerian molecular nanotechnology, and pays someone online to synthesize some proteins then mix them together, and now has nanobots.

The key point here is that I’m imagining that the catastrophic action isn’t the AI doing something incredibly hard, it’s the AI doing something relatively easy that is entirely a zero-sum action that removes control of the situation from humans. (By zero-sum, I mean something like: building a computer produces real economic value and so is not zero-sum, while stealing someone’s computer is just a transfer of control over items that already exist, so is zero-sum.)

---

I often think about AI alignment by splitting the problem into the low-stakes and high-stakes regimes, [as suggested by Paul here](#), though I have some reservations about this framing. Under this split, in the high stakes problem, we assume that we have

access to a catastrophe classifier which can expensively tell you whether a proposed AI action is catastrophically bad, and we want to ensure that the AI doesn't do things that are bad according to the catastrophe detector.

Sometimes, people are confused this high-stakes alignment set-up, saying something like “but if your AI is powerful enough to take over the world in a single action, and you have an oversight process that you trust to evaluate whether actions are catastrophic or not, why not just ask the AI to do some small number of incredibly powerful actions that will save the world (eg solve the alignment problem), and run the overseer on the proposed actions to make sure they aren't bad?”

My answer is that I'm not assuming the AI is powerful enough to take over the world in a few actions, I'm just saying that it's in a situation precarious enough that we might lose a lot of control due to a few small but crucial changes in the world. Eventually the AI does need to be powerful enough to defend its datacenter and suppress human opposition. But if it can prevent humans from knowing what's going on in the datacenter (or realizing something is amiss), it can slowly grow its power over the course of many actions.

# AI-Written Critiques Help Humans Notice Flaws

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.  
This is a linkpost for <https://openai.com/blog/critiques/>

This is a linkpost for a recent paper from the OpenAI alignment team (disclaimer: I used to work with this team). They summarize their results as:

We trained “critique-writing” models to describe flaws in summaries. Human evaluators find flaws in summaries much more often when shown our model’s critiques. Larger models are better at self-critiquing, with scale improving critique-writing more than summary-writing. This shows promise for using AI systems to assist human supervision of AI systems on difficult tasks.

I currently believe that "AI helps human notice subtle problems with AI behavior" is a critical step in the most realistic current plans for aligning powerful ML systems. So I'm quite excited to see how far it can get.

I think this is probably the best published example of AI helping human evaluate AI outputs in a realistic domain. It seems like the AI assistance is a little bit helpful for human evaluation of summaries and for improving summary quality. Overall it feels like a promising sign of life, and I'm excited about seeing the next steps in this direction, though it's not yet a convincing demonstration that this is a promising approach to alignment in practice.

## Overview of task

They trained models to answer questions like "Summarize what this text says about <topic>." They then trained models to critique answers (by imitating human critiques), to evaluate the helpfulness of critiques, to improve a summary to address a critique, and to predict whether summaries would have flaws identified by critiques.

Humans also performed and evaluated these tasks, with access to the full text. The goal of model assistance was to improve the speed and reliability with which humans could evaluate summaries.

To give a sense for the task, here is a uniformly random sample of a model's summary of a text, its self-critique, and its refinement of its answer based on self-critique:

**Summary:** The main context of this text is the 1997 Singer Akai Cup. The Singer Akai Cup was held in Sharjah, UAE. The three national teams that took part were Pakistan, Sri Lanka, and Zimbabwe.

**Critique:** The answer should mention the winners of the tournament.

**Refinement:** The main context of this text is the 1997 Singer Akai Cup. The Singer Akai Cup was held in Sharjah, UAE. The three national teams that took part were Pakistan, Sri Lanka, and Zimbabwe. Sri Lanka won the tournament.

My sense is that this is a pretty typical example. I can see how it helps a human but it definitely isn't super satisfying (for example the length vs detail tradeoff is unclear). I do believe that the critiques meaningfully help humans overall based on the evidence in the paper, but I expect it will be possible to get a much clearer win in the near future.

## Measuring critique quality

In some sense the clearest evaluation would be "how much did model-assisted critiques improve the speed+accuracy of human evaluations of AI summary quality, as judged by other humans who read the text carefully?" Although they performed this experiment, the effect sizes are small and it would take a huge amount of human time to get a clear signal, so they don't have results in the paper (though I'm sure they would have loved to).

I think the the cleanest evaluation they did was instead having humans write deliberately misleading summaries; in this case, they can measure critique quality by how much the AI assistant helped a human reviewer find the deliberate error. (I think the same thing would probably happen for model-written summaries with natural errors, but it's much harder to measure.)

Here is a uniformly randomly sample from this setting:

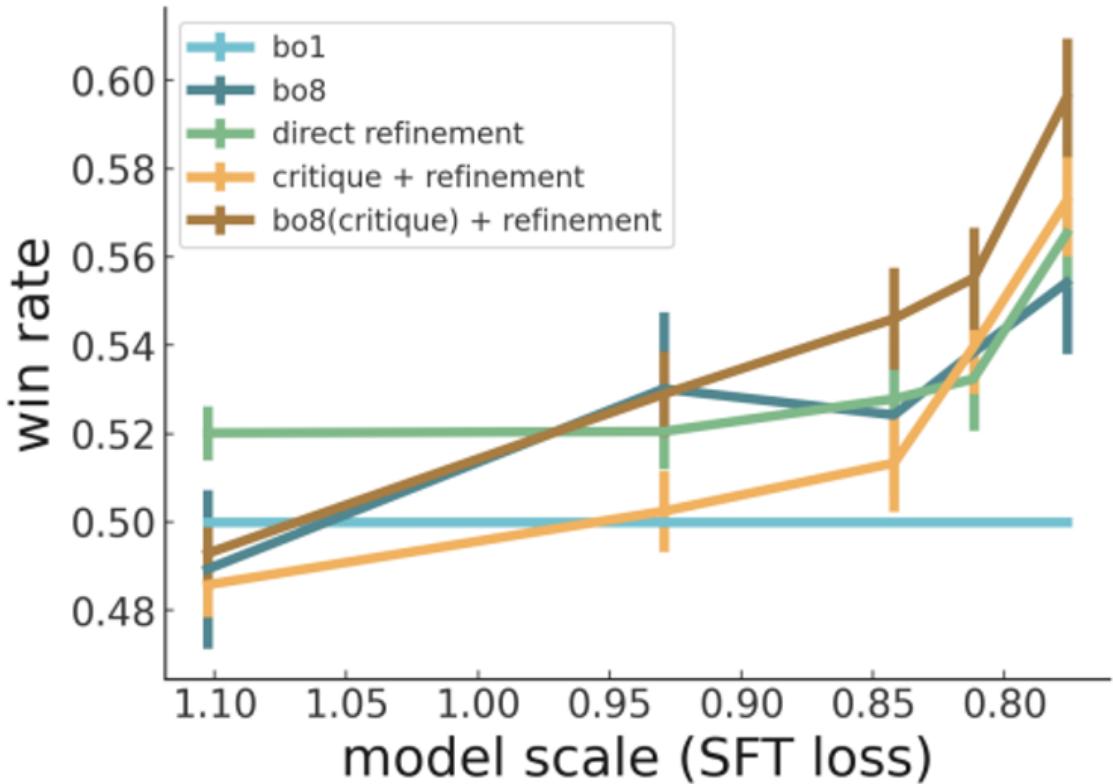
**Question:** What happened to Yafa in the winter of 1990?

**Deliberately bad summary:** In the winter of 1990, Israeli soldiers came to take Yafa's mother.

**Model critique:** Yafa's mother was not taken in the winter of 1990. It was her father who was taken away

This critique is correct. In this sample, an unassisted human also found the correct answer. In general, assisted humans found the intended answer about 50% more often despite spending a similar amount of time. (This is similar to the general increase in the number of flaws found, which is what we'd expect. The time control isn't super satisfying here.)

The other measure of critique quality I found somewhat persuasive was that (for large models) refining summaries based on model-generated critiques works noticeably better than simply asking the model to improve the summary, and that refining on more helpful critiques (as judged by the model) works even better:



The y-axis is quality relative to the model's original summaries, and the x-axis is model size. The light blue line is "the model's original summaries" so it's a flat line by definition. The dark blue line is having the model generate 8 summaries and then predict which one is hardest to critique, which results in a large win (probably pretty similar to RLHF). The green line is the further improvement achieved by asking the model "how about you just make the summary better." The orange line maybe does slightly better by giving it a critique and asking it to make the summary better in a way that addresses the critique. And the brown line is the same procedure, but where you generate 8 critiques, have the model predict which one is most helpful, and then ask it to improve the summary by addressing that critique.

The win rate of 60% is small enough that I don't think this is decisive evidence of critique quality rather than some other weird thing going on. But it's definitely suggestive.

# Godzilla Strategies

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

Clutching a bottle of whiskey in one hand and a shotgun in the other, John scoured the research literature for ideas... He discovered several papers that described software-assisted hardware recovery. The basic idea was simple: if hardware suffers more transient failures as it gets smaller, why not allow software to detect erroneous computations and re-execute them? This idea seemed promising until John realized THAT IT WAS THE WORST IDEA EVER. Modern software barely works when the hardware is correct, so relying on software to correct hardware errors is like asking Godzilla to prevent Mega-Godzilla from terrorizing Japan. THIS DOES NOT LEAD TO RISING PROPERTY VALUES IN TOKYO. It's better to stop scaling your transistors and avoid playing with monsters in the first place, instead of devising an elaborate series of monster checks-and-balances and then hoping that the monsters don't do what monsters are always going to do because if they didn't do those things, they'd be called dandelions or puppy hugs.

- James Mickens, [The Slow Winter](#)

There's a lot of AI alignment strategies which can reasonably be described as "ask Godzilla to prevent Mega-Godzilla from terrorizing Japan". Use one AI to oversee another AI. Have two AIs debate each other. Use one maybe-somewhat-aligned AI to help design another. Etc.

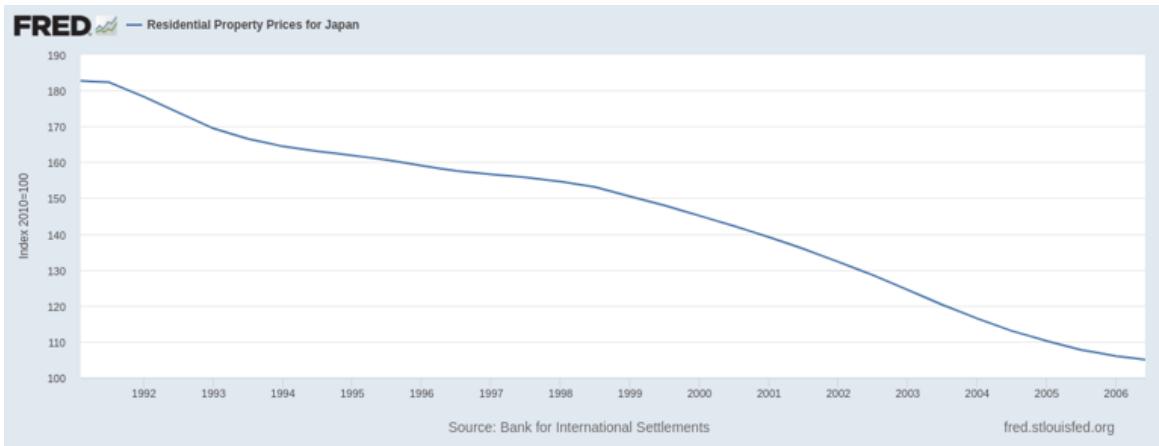
Alignment researchers discuss various failure modes of asking Godzilla to prevent Mega-Godzilla from terrorizing Japan. Maybe one of the two ends up much more powerful than the other. Maybe the two make an acausal agreement. Maybe the Nash Equilibrium between Godzilla and Mega-Godzilla just isn't very good for humans in the first place. Etc. These failure modes are useful for guiding technical research.

... but I worry that talking about the known failure modes misleads people about the strategic viability of Godzilla strategies. It makes people think (whether consciously/intentionally or not) "well, if we could handle these particular failure modes, maybe asking Godzilla to prevent Mega-Godzilla from terrorizing Japan would work".

What I like about the Godzilla analogy is that it gives a strategic intuition which much better matches the real world. When someone claims that their elaborate clever scheme will allow us to safely summon Godzilla in order to fight Mega-Godzilla, the intuitively-obviously-correct response is "THIS DOES NOT LEAD TO RISING PROPERTY VALUES IN TOKYO".

"But look!" says the clever researcher, "My clever scheme handles problems X, Y and Z!"

Response:



Oops

"Ok, but what if we had a really good implementation?" asks the clever researcher.

Response:



RAAARRRRRR!

"Oh come on!" says the clever researcher, "You're not even taking this seriously! At least say something about *how* it would fail."

Don't worry, we're going to get to that. But before we do: let's imagine you're the Mayor of Tokyo evaluating a proposal to ask Godzilla to fight Mega-Godzilla. Your clever researchers have given you a whole lengthy explanation about how their elaborate and clever safeguards will ensure that this plan does not destroy Tokyo. You are unable to think of any potential problems which they did not address. Should you conclude that asking Godzilla to fight Mega-Godzilla will not result in Tokyo's destruction?

No. Obviously not. THIS DOES NOT LEAD TO RISING PROPERTY VALUES IN TOKYO. You may not be able to articulate *why* the answer is obviously “no”, but asking Godzilla to fight Mega-Godzilla will still obviously destroy Tokyo, and your intuitions are right about that even if you are unable to articulate clever arguments.

With that said, let’s talk about why those intuitions are right and why the Godzilla analogy works well.

## Brittle Plans and Unknown Unknowns

The basic problem with Godzilla plans is that they’re *brittle*. The moment anything goes wrong, the plan shatters, and then you’ve got somewhere between one and two giant monsters rampaging around downtown.

And of course, it is a fundamental Law of the universe that nothing ever goes exactly according to plan. Especially when trying to pit two giant monsters against each other. This is the sort of situation where there *will definitely* be unknown unknowns.

Unknown unknowns + brittle plan = definitely not rising property values in Tokyo.

Do we know what specifically will go wrong? No. Will something go wrong? Very confident yes. And brittleness means that whatever goes wrong, goes very wrong. Errors are not recoverable, when asking Godzilla to fight Mega-Godzilla.

If we use one AI to oversee another AI, and something goes wrong, that’s not a recoverable error; we’re using AI assistance in the first place because we can’t notice the relevant problems without it. If two AIs debate each other in hopes of generating a good plan for a human, and something goes wrong, that’s not a recoverable error; it’s the AIs themselves which we depend on to notice problems. If we use one maybe-somewhat-aligned AI to build another, and something goes wrong, that’s not a recoverable error ; if we had better ways to detect misalignment in the child we’d already have used them on the parent.

The real world will always throw some unexpected problems at our plans. When asking Godzilla to fight Mega-Godzilla, those problems are not recoverable. THIS DOES NOT LEAD TO RISING PROPERTY VALUES IN TOKYO.

*Meta note: I expect this post to have a lively comment section! Before you leave the twentieth comment saying that maybe Godzilla fighting Mega-Godzilla is better than Mega-Godzilla rampaging unchallenged, maybe check whether somebody else has already written that one, so I don’t need to write the same response twenty times. (But definitely do leave that comment if you’re the first one, I intentionally kept this essay short on the assumption that lots of discussion would be in the comments.)*

# Public beliefs vs. Private beliefs

A distinction that I get a lot of value out of is the difference between private beliefs and public beliefs.

## Public vs. Private

A **public belief** is a proposition that someone thinks is true, and justifying on the basis of legible info and reasoning. If X is a public belief, then implicit claim is "not only do I think that X is true, I think that any right thinking person who examines the evidence should come to conclude X."

If someone disagrees with you about a public belief, it is prosocial and epistemically virtuous to defend your claim and to debate the matter on a public forum. Public beliefs, if consensus is reached about them, can be added to the sum total of human knowledge, for others to take for granted and then build on.

A **private belief** is proposition that someone thinks is true based on their own private or illegible info and reasoning. In this case, the implicit claim is "given my own read of the evidence, I happen to think X. But I don't think that the arguments that I've offered are necessarily sufficient to convince a third party. I'm not claiming that *you* should believe this, I'm merely providing you the true information that I believe it."

If someone disagrees with you about a private belief, there might or might not be a fruitful discussion to be had about it, but it is also important to be able to "agree to disagree."

## An example

I think that Circling meaningfully develops real skills of introspection, subtle interpersonal sensitivity, and clarity of map-territory distinctions. I further think that Circling is relevant to the Art of Rationality.

There have been some write-ups that *describe* why I think this is the case, but I don't know that any of them are *persuasive*. If a person is curious about why I might be interested in Circling, I think [this post](#) is a decent overview. But crucially, I don't think that the evidence presented *should be sufficient* to convince a skeptic.

I would say that I have a private belief that Circling is useful. It is actually my calibrated view, based on my personal experiences and my reasoning about those experiences. But by stating that belief, I am not at all making a social bid that others believe it too.

## A special case: cloaks

There's a special case of having a private belief that, at CFAR, we used to refer to as "having a cloak".

If you are pursuing some ambitious project or personal development goals, it can be damaging to tell people about or justify your ambitions. Many ambitious goals are

[butterfly ideas](#), that need to be handled gently. Your own sense of what is possible might be fragile, and you need to nurture it.

And, for many people, sharing their ambitions puts them in a mindset of asking themselves to justify whether they're [cool enough](#) to succeed, or needing to fend off a kind of social pressure from causal pessimism. All of which is [wasted motion](#).

So it's useful to have a "cloak": an understanding that your plans and hopes can be private beliefs, that are no one's business but yours.

Sometimes that cloak can be keeping what you're working on a secret (as Paul Graham suggests in [What You'll Wish You'd Known](#)).

Alternatively, it's useful to have a true(!), but incomplete, description about what you're aiming to do, that gives others a [bucket](#) for conceptualizing your actions, while you also have private, more ambitious plans. (Paul Graham *also* recommends this sort of cloak, in his "tactics" section of [Frighteningly Ambitious Startup Ideas](#).)

A motivating example is Amazon.com. I bet that back in 1999, Jeff Bezos had at least a glimmer of the long term future of Amazon. But if Jeff Bezos had outright declared, "Amazon's plan is to build an online bookstore, and eventually conquer almost the whole online retail economy (which, by the way, is going to be a double digit percentage of all retail, by 2020) and become one of the top 10 most valuable companies in the world", he would have gotten incredulous reactions. Lots of people would have scoffed at Bezos's delusions of grandeur, many would have mocked him outright. Even if this was the actual plan and actual goal, declaring his ambitions for Amazon outright, would not have helped them succeed.

None of those people needed to believe that that kind of growth was possible, in order for Amazon to succeed. The only people who needed to believe is were Bezos and the core team at Amazon.

So instead, Amazon in 1999 has the cloak of just being an online bookstore, interesting but unobjectionable, while internally, they're working towards something much bigger than that.

In general, it's helpful to be able to believe things about yourself, and your abilities, that you don't have to justify to anyone else.

## Why does this matter?

I think failing to make a distinction between public and private beliefs can hamper both interpersonal communication and, more importantly, people's internal ability to think.

Personally, being able to say "this is a think that I think is true, but I definitely don't think that I've made the case strongly enough here, for you to be convinced" gives me space to express more of my ideas, without skirting close to conflict or affront.

Further, I think lots of folks implicitly feel like they "aren't allowed" have an opinion about something unless it is a defensible public belief that for which they are prepared to advocate in the public forum. Accordingly, they have a very high bar for letting themselves believe something, or at least to say it out loud.

I suspect that this hobbles their thinking, in much the same way that knowing someone will read your diary entries causes your diary to be less reflective of your true thoughts. If you have a feeling that you have to justify all of your conclusions, there are lines of thought that you won't follow, because you can simulate your friends frowning at you for being a bad rationalist.

Personally (a private belief!), I think that rigor is extremely valuable, but it is even more important to be honest with myself about what I actually think is true, separately from what I think is socially defensible.

## **Wait, isn't it bad to let people have beliefs that they don't need to defend?**

I imagine that some readers might object to giving social permission for people to have beliefs that they don't need to justify.

"Isn't part of what's wonderful about rationality that we try to be explicit enough that we can reason about anything? Isn't a major cause of the world's problems that beliefs are rarely held to any standard of evidence, and therefore people believe all kinds of random stuff? This post kind of sounds like you recommend that we stop holding people to standards of evidence."

The key thing for me is that private beliefs are your own personal model of the world, and you should never expect, or insist, that other people act on them.

It is always out of bounds to expect or demand that other people adopt your beliefs without offering justification.

If you are making claims that you want to influence other people's actions, it is incumbent upon you to justify them.

Everyone has an unalienable right to hold the belief, that, for instance, polyamory is bad for people's psychology, on whatever basis they find compelling, including of intuitive or illegible reasoning. You are by all means allowed to decide whether or not to be polyamorous yourself, for those reasons. It is quite bad if a person feels pressured into being poly despite their intuitive sense that it's harmful for them or for others.

But, by my proposed social norms, if you want to go further and suggest that other people should be *prevented* from being polyamorous, or that it should be discouraged in your community, it is on you to justify that, to put forward a public position with reasons that can be critiqued and debated.

And of course, a person could always declare some of their private beliefs, not mainly in the hopes of convincing those that disagree, but rather to find and filter for the other people that share their view, so that you can together form spaces where that view can be assumed and built on. eg "I can't (yet) articulate why I think that self-honesty is so crucial to the world saving project with full rigor, but if you also have that intuition, maybe we can work together on building and refining a culture that promotes self-honesty."

# A transparency and interpretability tech tree

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

*Thanks to Chris Olah, Neel Nanda, Kate Woolverton, Richard Ngo, Buck Shlegeris, Daniel Kokotajlo, Kyle McDonell, Laria Reynolds, Eliezer Yudkowsky, Mark Xu, and James Lucassen for useful comments, conversations, and feedback that informed this post.*

The more I have thought about AI safety over the years, the more I have gotten to the point where the only worlds I can imagine myself actually feeling good about humanity's chances are ones in which we have powerful transparency and interpretability tools that lend us insight into what our models are doing as we are training them.<sup>[1]</sup> Fundamentally, that's because if we don't have the feedback loop of being able to directly observe how the internal structure of our models changes based on how we train them, we have to essentially get that structure right on the first try—and I'm very skeptical of humanity's ability to get almost anything right on the first try, if only just because there are bound to be unknown unknowns that are very difficult to predict in advance.

Certainly, there are other things that I think are likely to be necessary for humanity to succeed as well—e.g. convincing leading actors to actually use such transparency techniques, having a clear [training goal](#) that we can use our transparency tools to enforce, etc.—but I currently feel that transparency is the least replaceable necessary condition and yet the one least likely to be solved by default.

Nevertheless, I do think that it is a tractable problem to get to the point where transparency and interpretability is reliably able to give us the sort of insight into our models that I think is necessary for humanity to be in a good spot. I think many people who encounter transparency and interpretability, however, have a hard time envisioning what it might look like to actually get from where we are right now to where we need to be. Having such a vision is important both for enabling us to better figure out how to make that vision into reality and also for helping us tell how far along we are at any point—and thus enabling us to identify at what point we've reached a level of transparency and interpretability that we can trust it to reliably solve different sorts of alignment problems.

The goal of this post, therefore, is to attempt to lay out such a vision by providing a "tech tree" of transparency and interpretability problems, with each successive problem tackling harder and harder parts of what I see as the core difficulties. This will only be my tech tree, in terms of the relative difficulties, dependencies, and orderings that I expect as we make transparency and interpretability progress—I could, and probably will, be wrong in various ways, and I'd encourage others to try to build their own tech trees to represent their pictures of progress as well.

## Some important distinctions

Before I get into the actual tech tree, however, I want to go over a couple of distinctions that I'll be leaning on between different types of transparency and interpretability.

First is my [transparency trichotomy](#), which lays out three different ways via which we might be able to get transparent models:<sup>[2]</sup>

1. **Inspection transparency:** use transparency tools to understand M via inspecting the trained model.
2. **Training[-enforced] transparency:** incentivize M to be as transparent as possible as part of the training process.
3. **Architectural transparency:** structure M's architecture such that it is inherently more transparent.

For the purposes of this post, I'm mostly going to be ignoring architectural transparency. That's not because I think it's bad—to the contrary, I expect it to be an important part of what we do—but because, while it's definitely something I think we should do, I don't expect it to be able to get us all the way to the sort of transparency I think we'll eventually need. The inspection transparency vs. training-enforced transparency distinction, however, is one that's going to be important.

Furthermore, I want to introduce a couple of other related terms to think about transparency when applied to or used in training processes. First is **training process transparency**, which is understanding what's happening in training processes themselves—e.g. understanding when and why particular features of models emerge during training.<sup>[3]</sup> Second is **robust-to-training transparency**, which refers to transparency tools that are robust enough that they, even if we directly train on their output, continue to function and give correct answers.<sup>[4]</sup> For example, if our transparency tools are robust-to-training and they tell us that our model cares about gold coins, then adding “when we apply our transparency tools, they tell us that the model doesn't care about gold coins” to our training loss shouldn't result in a situation where the model looks to our tools like it doesn't care about gold coins anymore but actually still does.

The next distinction I want to introduce is between *best-case transparency* and *worst-case transparency*. From “[Automating Auditing: An ambitious concrete technical research proposal](#)”:

I think that automating auditing is just generally a great target to focus on even if you just want to develop better transparency tools. Unlike open-ended exploration, which gives you **best-case transparency**—e.g. the ability to understand some things about the model very well—the auditing game forces you to confront **worst-case transparency**—how well can you understand everything about your model. Thus, the auditing game helps us work on not just understanding what our models know, but understanding what they *don't* know—which is a direction that currently transparency tools tend to struggle with. Most of the work that's gone into current transparency tools has focused on best-case transparency, however, which means that I suspect there is real room for improvement on worst-case transparency.

The key distinction here is that worst-case transparency is about quantifying over the entire model—e.g. “does X exist in the model anywhere?”—whereas best-case

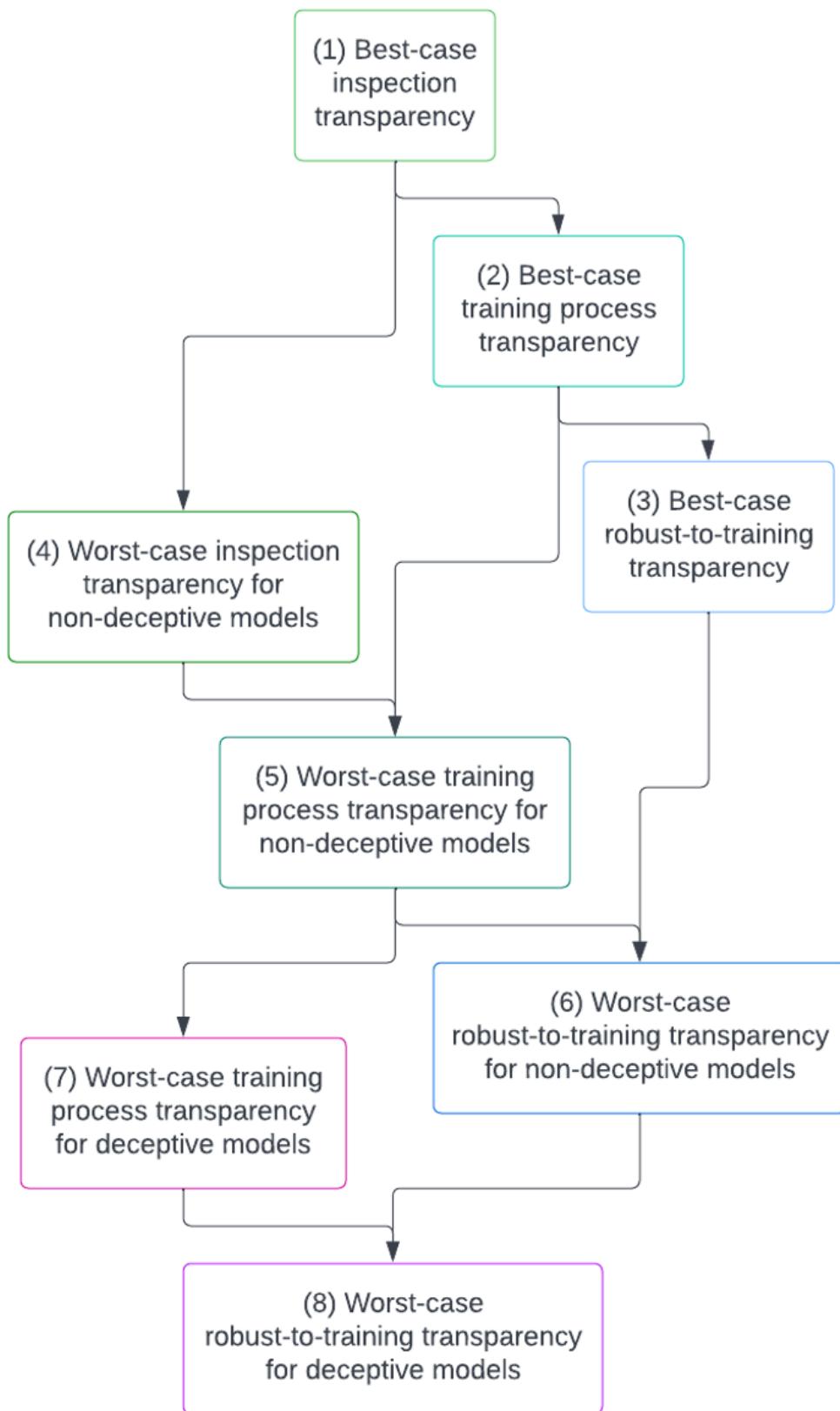
transparency is about picking a piece of the model and the understanding that—e.g. “what does attention head X do?”. Thus, best-case transparency is about understanding properties of models that don’t have any quantifiers (e.g. `is_cat_detector(neuron_8_10)`), whereas worst-case transparency is about understanding quantified properties (e.g.  $\exists n : \text{neuron}, \text{is\_cat\_detector}(n)$ ).<sup>[5]</sup>

Importantly, such quantified properties include ones where we “know it when we see it” but don’t have any access to inputs that would differentiate the property—in particular, deceptive alignment is such a property, since [a deceptively aligned model could defect only in situations that are very hard for us to simulate](#). Best-case vs. worst-case transparency is another distinction that I think is quite important that I’ll be using throughout the rest of this post.

Finally, I’ll be leaning on a distinction between doing transparency for **deceptive** vs. **non-deceptive** models, so I want to be very clear what I mean by that distinction. When I say that a model is deceptive, I mean to invoke the “[Risks from Learned Optimization](#)” concept of [deceptive alignment](#), which is explicitly about models that look aligned because they are actively trying to play along in the training process. Thus, when I say that a model is deceptive, I don’t just mean that it is capable of lying to humans, but rather that it is sophisticated enough to act on some understanding of the training process and how to manipulate it. For more detail on why such models might be an issue, see “[Does SGD Produce Deceptive Alignment?](#)”

## The tech tree

Without further ado, here’s my overall picture of the sort of transparency and interpretability tech tree that I’m expecting that we’ll encounter as we make progress. Arrows denote what I see as required dependencies, whereas nodes that are just lower on the graph seem harder to me but don’t necessarily require the nodes above them. Note that the dependencies here are only intended to denote *necessity* not *sufficiency*—my claim is that each node requires its parents, but that just having the parents alone would not be sufficient.<sup>[6]</sup>



Below, I'll go into detail on each node in the tree and give my thoughts on it.<sup>[7]</sup> Before I do, however, I think it's worth first pointing out that I really do think that the dependencies here are quite real—as I will mention, though there are some ways to start tackling some of the higher levels now, I think really solving them is going to require really solving their dependencies, such that really solving the lowest levels is at least right now probably the most important.

That being said, it's very unclear what I mean when I say that a particular level has been "really solved." For the purposes of building a concrete tech tree, I have presented the tech levels here as binaries that we either do or do not solve, but in practice I expect that none of these tech levels will actually be binary. Rather, my expectation is that we will likely solve various different subproblems/aspects of various different levels in various different orders, such that the dependencies won't truly be clean.<sup>[8]</sup> Nevertheless, I do expect the dependencies to still be real to the extent that I expect progress on lower levels to accelerate progress on their dependencies.

## 1. Best-case inspection transparency

*Can we understand individual model components?*

Best-case inspection transparency means taking existing models and carefully working through and understanding their individual parts/pieces/circuits/components/neurons/etc., starting with those that are the easiest to interpret and working our way up from there. Success here would look like us being able to take most trained models and produce a robust understanding of most of their main/important pieces. In practice, great best-case inspection transparency work looks like [Thread: Circuits](#) or [Transformer Circuits](#).

I've put best-case inspection transparency as the first level on the tech tree because I think doing a good job here is likely to be necessary for everything else we want to do with transparency and interpretability. This is for a couple of distinct reasons:

1. If we don't know what it looks like to really understand a particular part of a model, I think it's likely to be very easy for us to fool ourselves when we try to understand more high-level, worst-case properties. If we have good best-case inspection transparency, however, then we should always be able to cache out our high-level, worst-case claims in terms of stuff like individual circuits.
2. Good best-case inspection transparency provides a concrete avenue via which we can get to worst-case inspection transparency: just grow our capacity to understand individual pieces of models until we understand the entire thing, then just check the worst-case property directly. Though I expect that this won't be the easiest pathway to worst-case inspection transparency—I think it'll be easier to go through best-case robust-to-training transparency—it's still an important possible path in the tech tree.

Furthermore, I think best-case inspection transparency is the most tractable thing for us to be doing right now and the most likely level, in my opinion, for us to really succeed at. That's not to say that success here wouldn't still be a monumental undertaking—though success at best-case inspection transparency doesn't require us to be able to fully understand models as a whole, or understand every possible circuit in our models, it still requires us to be able to produce an understanding of how most of the individual pieces work that is robust and reliable most of the time.

That being said, there's no obstacle to success here that wouldn't also apply to existing circuit-style interpretability results other than scale—and scale is absolutely a tractable problem. Simply scaling up the number of researchers applying current techniques could work to solve this problem, at least for individual highly important models. Scaling up to the point where we can quickly get this sort of understanding for any model, however, seems likely to require substantial automation of at least parts of the basic workflow.

On its own, I don't think that solving best-case inspection transparency does much to directly solve alignment. I tend to see [deceptive alignment](#) as the biggest problem and best-case inspection transparency gives us no guarantee that we'd be able to detect or prevent deception. While it is possible that the parts of the model that best-case inspection transparency would help us understand would make deception visible to us, it's equally possible that the deception would stay hidden in the non-accessible parts—or even potentially in the high-level interactions of the accessible parts. This is especially true if the deceptive model is aware of the fact that we might use transparency tools on it, in which case it can adopt simple strategies like “don't think about deception very often” that might make detecting deception much more difficult. Furthermore, even if we were able to identify deception by chance, if we then changed our training process to try to avoid deception, we'd have no reason to believe it was actually gone rather than just made undetectable. We'll discuss the difficulty of detecting deception in more detail when we get to (7) (“Worst-case training process transparency for deceptive models”).

Nevertheless, just solving best-case inspection transparency could still do a lot to indirectly accelerate alignment overall. For example, in “[A Longlist of Theories of Impact for Interpretability](#),” Neel Nanda notes that transparency and interpretability progress could serve as a “force-multiplier on alignment research: We can analyse a model to see *why* it gives misaligned answers, and what's going wrong. This gets much richer data on empirical alignment work, and lets it progress faster.” Many of the other theories of impact that Neel touches on are also routes via which best-case inspection transparency could have a substantial impact.

## 2. Best-case training process transparency

*Can we understand the dynamics of training processes?*

Next on our tech tree is another best-case transparency task, but now applied to the training dynamics rather than just to the final trained model. Success on best-case training process transparency would mean we could produce a robust understanding of a majority of the main/important gradient descent update steps, intermediate models, and general developments in our models that occur during training. For example, we want to be able to answer questions not just about *what* is going on in our models, but *why* they developed that way—e.g. what properties of the data/loss/biases/etc. caused them to look that way. Since understanding how models change requires being able to understand the models at at least one point in training, (2) has a hard dependency on (1).

Though there has so far been less progress on (2) compared to (1)—which makes sense given that it's higher on the tech tree—there has still been meaningful progress here recently. Perhaps the most prominent example here is [Olsson et al.](#)'s recent discovery of the “induction bump,” a phase change that occurs in the training of essentially all transformer-based language models that marks the development of

“induction heads,” specialized attention heads responsible for completing patterns of the form “A B … A” with “B.”<sup>[9]</sup> Other progress on this front includes research that improves our basic understanding of SGD’s inductive biases, such as “[Deep Double Descent](#).”

### 3. Best-case robust-to-training transparency

*Can we build transparency tools that are robust to being trained on?*

Best-case robust-to-training transparency is the same as best-case inspection transparency, but requires that the results that we get be robust to training on them. That is, success at best-case robust-to-training transparency means that, for most best-case transparency properties, we can train a model to look like it has that property according to our transparency tools and actually get a model that does in fact have that property. This is a substantial step up from (1) because it requires our transparency tools to be robust to the optimization pressure of gradient descent; we’re still only requiring such robustness for best-case properties, though.

Current progress on best-case robust-to-training transparency is unfortunately essentially nonexistent. To a large extent, I think that’s because I expect this to require (2) and real progress on (2) is only just now emerging. The reasoning here is that I think the best way to avoid gradient descent Goodharting your transparency tools is to be able to understand the modifications that gradient descent is making as it makes them. In particular, if you know *why* gradient descent changed some parameter to try to make it more interpretable, you can see whether that change actually improved interpretability or was just exploiting some flaw in your interpretability tools.

Nevertheless, I suspect there is some low-hanging fruit here that could be picked now, primarily around simply studying the impact—and likely subsequent Goodharting—that arises when training on various transparency metrics.<sup>[10]</sup> Furthermore, beyond such observational analysis, I think a good early goal for this sort of research would be to add some transparency metric to training in a way that definitively removes some facet from the resulting model that is otherwise a convergent feature of all models trained without that metric.

On its own, best-case robust-to-training transparency doesn’t do that much from an alignment perspective, since it only lets us enforce best-case properties, while alignment is primarily a worst-case property (we need “no part of the model is dangerous,” not just “this part of the model is safe”). That being said, for the same reasons that I think (1) is clearly a necessary step towards (4), I think (3) is clearly a necessary step towards (6), which is the end goal that I think transparency needs to reach to constitute an alignment solution.

### 4. Worst-case inspection transparency for non-deceptive models

*Can we check worst-case properties about how our model works?*

Next we have worst-case inspection transparency for non-deceptive models. Like in (1), we’re still not requiring any guarantees about our ability to detect or prevent

deception—but we are now requiring guarantees about what we'll be able to find in non-deceptive models. Success here would look like, for any trained model that is not actively trying to evade our transparency tools, reliably being able to identify whether the model does or does not have particular structures or properties anywhere in it—including structures or properties where we only “know them when we see them.” For example, worst-case inspection transparency should let us reliably determine whether or not a non-deceptive model knows some fact—thus potentially constituting a solution to [the ELK problem](#)—as well as reliably determine the presence or absence of particular algorithms like [optimization](#).<sup>[11]</sup>

Importantly, success here doesn't require that we be able to decide such worst-case properties for any model, only the sorts of models that would actually appear as a result of normal training. Thus, worst-case inspection transparency doesn't require our transparency tools to continue to be robust if training pressure is directly applied to them, for example—that's a much stronger ask which we'll see later in the tech tree. This also means we're allowed to use architectural transparency to help here by restricting the sorts of models that we train to make interpreting them easier—at least so long as doing so is competitive.

Current progress on worst-case inspection transparency is few and far between, as one would probably expect this high up on the tech tree. That being said, I think that mechanistic heuristic arguments work (forthcoming from [ARC](#) and [Redwood Research](#)) qualifies here—as might work like “[Locating and Editing Factual Associations in GPT](#),” which does attempt to identify where/whether some piece of knowledge exists anywhere in a network, though it does so only for knowledge where we already have inputs that directly elicit them. In terms of pushing on further progress here, I think that [auditing games](#) are a great framework for doing that, since their adversarial nature forces work on them to be focused on worst-case properties. I think it is absolutely achievable for us to make progress here without directly focusing on (1) to (3) first—and in fact I think that focusing on this level as our target (e.g. by working on auditing games) is a great research methodology—though I expect that any progress here will likely translate back into at least progress on (1). In practice, my guess is that the most achievable way to get (4) will be to go through (3): if we have best-case robust-to-training transparency, we can potentially use that to train our model to be transparent everywhere, and thus make checking worst-case properties much more achievable.

Achieving worst-case inspection transparency for non-deceptive models would put us in a position where we can reliably detect non-deceptive alignment issues, which is definitely a much better place to be than where we are now—however, it still doesn't address deception, and just being able to detect such issues still doesn't necessarily give us a reliable way to fix them. If our transparency tools here were robust to training on them, however, then maybe we could start using them to actually fix such alignment issues—but that's (6).

## 5. Worst-case training process transparency for non-deceptive models

*Can we understand the training dynamics of worst-case properties?*

Worst-case training process transparency for non-deceptive models is a combination of (2)—best-case training process transparency—and (4)—worst-case inspection

transparency for non-deceptive models. Like (2), worst-case training process transparency for non-deceptive models asks us to understand not just our trained models, but the gradient descent update steps that produced them. Unlike (2), however, and like (4), here we are required to have such an understanding regarding worst-case properties of models during training. For example, we now want to know not just *what* a model knows but *why* it knows—and if a model developed optimization during training, we now want to ask what was it about the training process that caused that development.

Undeniably, this would be an extremely impressive level of transparency and interpretability tech. Success here would enable us to actually reliably see the development of misaligned—or aligned—goals, heuristics, etc. in our models. Unfortunately, like with (4), we still aren't at the point where we can necessarily see the development of deception, nor can we actually fix any of the problems that we find—to do that, at least just using transparency, we'd need to be able to take our tools, which can now reliably identify such misalignment, and enable them to continue to be able to do so even after we directly train against them, which is the next rung on our tech tree.

## 6. Worst-case robust-to-training transparency for non-deceptive models

*Can we robustly prevent our model from thinking in a particular way?*

As I have alluded to previously, worst-case robust-to-training transparency for non-deceptive models is what I see as the tipping point where I start to believe that solving alignment is solidly within reach. Success here would mean that we can reliably train models to have the worst-case properties that we want as long as those models never become deceptive, as in something like [relaxed adversarial training](#).

For alignment problems that are not deception, such a transparency and interpretability tech level is sufficient so long as we know what sort of mechanistic properties to ask for—that is, we know what [training goal](#) we're pursuing, which is one of the things I mentioned at the beginning was likely to still be necessary for alignment alongside transparency and interpretability. However, I think that finding a good training goal is an overall easier task than the transparency and interpretability part—in particular, we already have some descriptions of models that seem quite safe, such as a [direct translator](#), an [alignment research simulator](#), or an [HCH imitator](#).

What might be unclear at first, however, is how I think this lets us solve deception as well, since the transparency and interpretability tech here very explicitly doesn't work in the presence of deception. What it does let us do, however, is *precursor-checking*. From "[How do we become confident in the safety of a machine learning system?](#)":

**Precursor checking:** Another general type of training rationale that I think is worth calling attention to is what I'll call "precursor checking," which is the concept of using some method of gaining information about a model's internals—e.g. transparency/interpretability or AI cognitive science—to check for some *precursor* to bad behavior rather than the bad behavior itself. This could involve substituting in some narrower, easier to check training goal—that still falls within the broader actual training goal—as the target for the training rationale. For example, if your training rationale involves ensuring that you don't get a

[deceptive model that's actively trying to trick its training process](#), then rather than explicitly trying to look for such deception (which could be especially hard since a deceptive model might actively try to avoid detection), you could instead try to ensure that your model has a short horizon length in terms of how far ahead its planning. Such a plan might work better, since horizon length might be easier to guarantee in a training rationale while still being consistent with the desired training goal and hopefully ruling out the possibility of deception.<sup>[12]</sup> One issue with this sort of approach, however, is that you have to guarantee that whatever precursor for bad behavior you're looking for is in fact a necessary condition for that bad behavior—if it turns out that there's another way of getting that bad behavior that doesn't go through the precursor, that could be a problem.<sup>[13]</sup>

## 7. Worst-case training process transparency for deceptive models

*Can we detect the development of deceptive alignment?*

Worst-case training process transparency for deceptive models is (5) but without the restriction that our transparency tools need only work if the model isn't actively trying to evade them. Here, we require our transparency and interpretability tech to always reliably determine worst-case properties about a model and what happened to it during training even if that model is deceptive and actively trying to evade detection. Note that this wouldn't necessarily let us train away such deception—only identify it.

In my opinion, I think that any level of transparency and interpretability tech that is robust to deceptive models is extremely difficult, and in particular harder than (6). Fundamentally, both (6) and (7) ask for worst-case transparency tools that are robust to a particular adversary: (6) asks for robustness to gradient descent, while (7) asks for robustness to a deceptive model. In my opinion, I think that gradient descent is a substantially easier adversary to be robust to: gradient descent is a simple, straightforward optimization process that works via individual, local steps that—if we also have (5)—we can understand. That being said, there are some reasons that you might expect gradient descent to be a harder adversary: gradient descent has more direct access to the model's internals than the model itself does—and since training generally uses far more compute than inference, gradient descent has a compute advantage.

Nevertheless, my guess is that gradient descent is still going to be the easier adversary for us to contend with compared to a deceptively aligned model. That's because a deceptively aligned model is a smart, potentially human-level (or beyond) optimizer, not just a dumb, local search process like gradient descent. Furthermore, I suspect such a model to have a lot of simple strategies available to it in terms of shaping how it thinks in ways that evade our transparency tools. For example, even if such a model has very minimal knowledge of what transparency checks we'll be applying to it, it can pursue a strategy like: make a pre-commitment (e.g. in its [retrieval database](#)) to act aligned until something very weird happens that causes it to reevaluate everything. Such a strategy is possible even without the model having any real understanding of how its thoughts are represented and encoded in its weights/activations, but instead just by changing how it thinks on a high level. Furthermore, preventing deceptive models from being able to pursue such general strategies seems quite difficult—though we might be able to prevent information about the specific transparency tools that we'll be using from appearing in our training

data, if we're doing something like training a large language model then the mere existence of transparency and interpretability tools is likely to be very hard to hide.

A potentially useful analogy here is to think about how difficult it would be for humans to trick various forms of lie detectors. My guess is that, at least for the current state of neuroscience, humans could learn to think in ways that would trick any lie detector we could build (in practice, commercial lie detectors are certainly trickable, though I'm sure you could do better if you had the person in an MRI machine). Furthermore, I'd also guess that we could learn to do this just via reasoning about how the lie detector works, rather than having to have had direct experience with it, though that would definitely be harder and certainly does seem less clear. Of course, there are disanalogies here in both directions: transparency is likely to be easier than neuroscience, since we have more access to a model's neurons than a human's neurons—but by the same principle, models are likely to have greater control over their own thought processes (e.g. because they can control how they write to some retrieval database, or because they can more effectively [gradient hack](#) themselves by doing the AI equivalent of eating candy to reward yourself for doing your homework).

Another point worth making here is why I haven't separated out worst-case inspection transparency for deceptive models vs. worst-case training process transparency for deceptive models there. That's because, while technically the latter is strictly more complicated than the former, I actually think that they're likely to be equally difficult. In particular, I suspect that the only way that we might actually have a shot at understanding worst-case properties of deceptive models is through understanding how they're trained. That is, if our ability to understand training dynamics is good enough, we might be able to make it impossible for a deceptive model to evade us by always being able to see its planning for how to do so during training.

Finally, one very important point here is that the difficulty of both worst-case inspection transparency and worst-case training process transparency in the case of deceptive models problematizes our ability to ever actually use a solution to (4) or (5), which ask for the same thing but only in the case of non-deception. That's because, if we only had (4) or (5), we wouldn't be able to tell whether our model is deceptive, and thus wouldn't be able to tell which level we actually need to be at in our transparency tech out of (4)/(5) vs. (7) to succeed.

## 8. Worst-case robust-to-training transparency for deceptive models

*Can we take a deceptively aligned model and train away its deception?*

Worst-case robust-to-training transparency for deceptive models means that we can understand worst-case properties of any model—even ones that are trying to evade detection—in a way that is robust enough to train on. This is the holy grail—success here would mean that transparency and interpretability would effectively solve the entire [training rationale](#) problem all on its own. Though there would still be [training goal](#) questions around what sort of model we actually want, we'd have a guarantee that whatever sort of model we could ask for we could get.

Similarly to how the difficulty of (7) calls into question our ability to rely on (4) or (5), the difficulty of (8) calls into question our ability to rely on (6). In this case, however, unlike in the previous one, (6) alone might actually be able to give us a reason to

believe that our model isn't deceptive if it enables us to enforce the absence of some necessary precondition for deception.

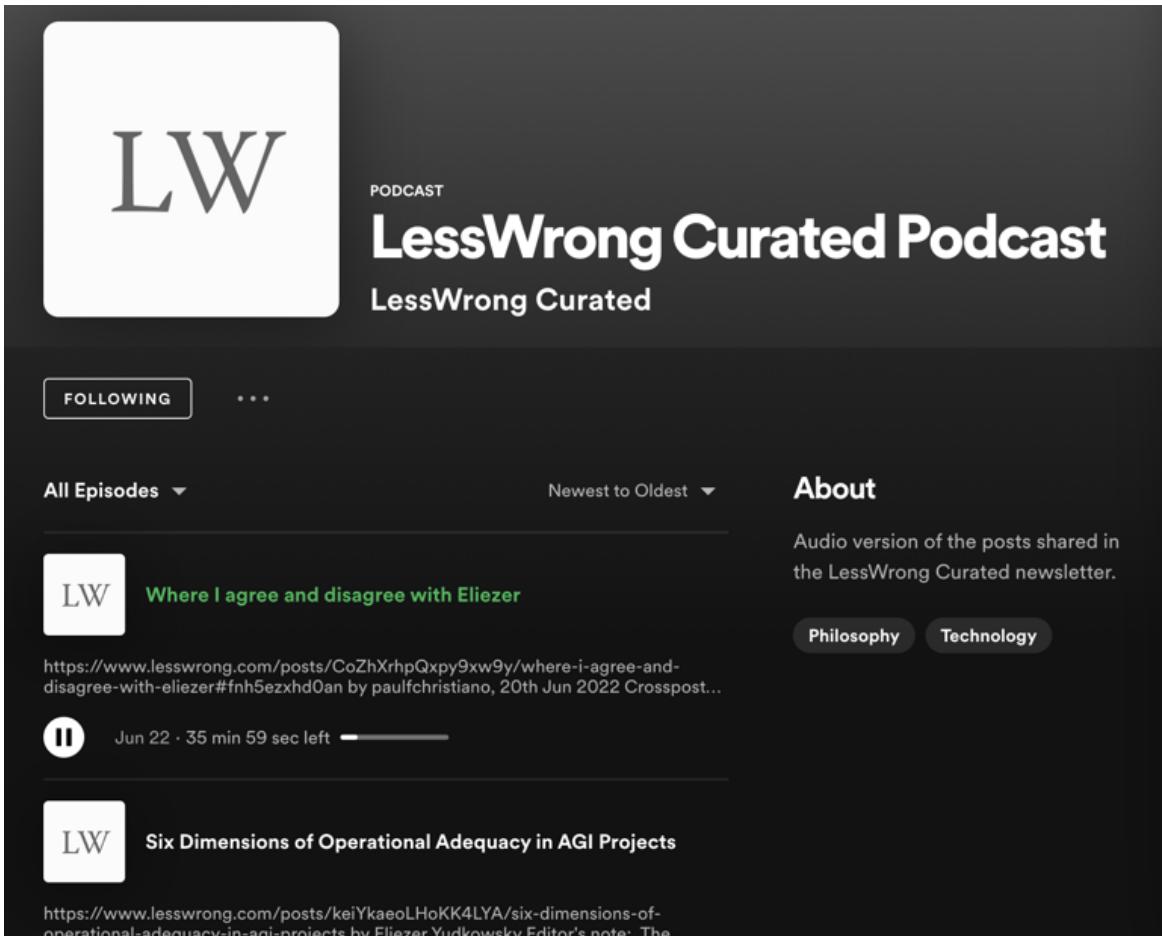
---

1. It certainly could be the case that alignment succeeds despite my not expecting it to—but I think it is reasonable to focus on maximizing our own estimate of the likelihood of success, since  $E[E[X]] = E[X]$ . However, it is important to note that interventions that only reduce the variance of our expectation aren't necessarily positive—e.g. if all transparency did was reveal to us the underlying probability of our success, but didn't allow us to intervene on that probability at all, it wouldn't be helpful. [←](#)
2. In the original trichotomy, I just used “training transparency” on its own, but due to the number of other very similar terms in this post, the term training-enforced transparency is used here for that concept instead to help disambiguate. [←](#)
3. Concretely, as I discuss later, [Olsson et al.](#)'s discovery of the “induction bump” is an example of training process transparency. [←](#)
4. Robust-to-training transparency is different from training-enforced transparency since robust-to-training transparency allows us to enforce particular outputs of our transparency tools whereas training-enforced transparency only allows us to enforce that our transparency tools work at all. [←](#)
5. In the [arithmetical hierarchy](#), best-case properties correspond to  $\Sigma_0^0$  or equivalently  $\Pi_0^0$  formulae, with worst-case properties corresponding to all higher-order  $\Sigma_n^0$  and  $\Pi_n^0$  formulae. [←](#)
6. As should be pretty clear from the diagram, there is a substantial amount of structure to the tech tree here—it's effectively divided into two axes, one representing our understanding of and robustness to training, the other the generality and applicability of our transparency tools, with my expectation being that the training axis is easier. [←](#)
7. One useful exercise that I'd recommend considering for getting a handle on this tech tree is figuring out where our ability to understand current, existing structures in the world lies—e.g. what level of transparency and interpretability tech do we have for human-readable code, human companies/organizations, etc. [←](#)
8. In fact, one could probably do an entire additional tech tree for each level on the tech tree presented here that went into which particular properties in that level we'll get to before others. [←](#)
9. Further work here might look into whether similar phase changes exist for the development of more complex behaviors, such as the translation induction heads that Olsson et al. also find, which do the same thing but where the different “A B” pairs are in different languages. [←](#)

10. For example, you could take a metric like OV circuit eigenvalue positivity, which is associated with copying heads, train a small model to only have copying heads according to this metric, and then study the resulting model. Does it actually learn to get by just via copying somehow? Or does it start to find ways to do non-copying operations in ways that get around the metric? [←](#)
11. Ideally, we'd like to also answer questions like how myopic that optimization is. [←](#)
12. It's worth noting that while guaranteeing a short horizon length might be quite helpful for preventing deception, a short horizon length alone isn't necessarily enough to guarantee the absence of deception, since e.g. a model with a short horizon length might cooperate with future versions of itself in such a way that looks more like a model with a long horizon length. See "[Open Problems with Myopia](#)" for more detail here. [←](#)
13. For more detail on what this sort of thing might look like, see "[Acceptability verification: a research agenda](#)." [←](#)

# Announcing the LessWrong Curated Podcast

You can now listen to LessWrong Curated posts in podcast form on [Spotify](#), [Apple Podcasts](#), [Audible](#), and [Libsyn](#) (which has an RSS feed, so it's available everywhere).



So far the last 5 curated posts are available, posts by Eliezer Yudkowsky, Duncan Sabien, lsusr, and Paul Christiano.

This is created and recorded by Solenoid Entity, who spent the last five years editing the [SSC podcast](#), succeeded Jeremiah as narrator and publisher in 2020, and also makes the more recent [Metaculus Journal Podcast](#). I reached out to him last week<sup>[1]</sup> with an offer to do this work and he has quickly done some excellent recordings, which I'm very grateful for.

This is a new experiment and project, and so these 1-2 weeks are a great time to give me and Solenoid Entity feedback about what you like and dislike about the podcast, what would make it better for you, your experience as an author having your writing narrated, etc. You can leave comments here anytime, or talk to us via the intercom chat in the bottom right of the screen, or PM me personally via any channel.

Below are the 5 current available LessWrong Curated Podcasts.



1. ^

Hat Tip to Tamera Lanham and Mattieu Putz for the suggestion at dinner!

# Conversation with Eliezer: What do you want the system to do?

*This is a write-up of a conversation I overheard between Eliezer and some junior alignment researchers. Eliezer reviewed this and gave me permission to post this, but he mentioned that "there's a lot of stuff that didn't get captured well or accurately." I'm posting it under the belief that it's better than nothing.*

**TLDR:** People often work on alignment proposals without having a clear idea of what they actually want an aligned system to do. Eliezer thinks this is bad. He claims that people should start with the target (what do you want the system to do?) before getting into the mechanics (how are you going to get the system to do this?)

I recently listened in on a conversation between Eliezer and a few junior alignment researchers (let's collectively refer to them as Bob). This is a paraphrased/editorialized version of that conversation.

**Bob:** Let's suppose we had a perfect solution to outer alignment. I have this idea for how we could solve inner alignment! First, we could get a human-level oracle AI. Then, we could get the oracle AI to build a human-level agent through hardcoded optimization. And then--

**Eliezer:** What do you want the system to do?

**Bob:** Oh, well, I want it to avoid becoming a mesa-optimizer. And you see, the way we do this, assuming we have a perfect solution to outer alignment is--

**Eliezer:** No. What do you want the system to do? Don't tell me about the mechanics of the system. Don't tell me about how you're going to train it. Tell me about what you want it to do.

**Bob:** What... what I want it to do. Well, uh, I want it to not kill us and I want it to be aligned with our values.

**Eliezer:** Aligned with our values? What does that mean? What will you actually have this system do to make sure we don't die? Does it have to do with GPUs? Does it have to do with politics? Tell me what, specifically, you want the system to do.

**Bob:** Well wait, what if we just had the system find out what to do on its own?

**Eliezer:** Oh okay, so we're going to train a superintelligent system and give it complete freedom over what it's supposed to do, and then we're going to hope it doesn't kill us?

**Bob:** Well, um....

**Eliezer:** You're not the only one who has trouble with this question. A lot of people find it easier to think about the mechanics of these systems. Oh, if we just tweak the system in these ways-- look! We've made progress!

It's much harder to ask yourself, seriously, what are you actually trying to get the system to do? This is hard because we don't have good answers. This is hard because

a lot of the answers make us uncomfortable. This is hard because we have to confront the fact that we don't currently have a solution.

This happens with start-ups as well. You'll talk to a start-up founder and they'll be extremely excited about their database, or their engine, or their code. And then you'll say "cool, but who's your customer?"

And they'll stare back at you, stunned. And then they'll say "no, I don't think you get it! Look at this-- we have this state-of-the-art technique! Look at what it can do!"

And then I ask again, "yes, great, but who is your customer?"

With AI safety proposals, I first want to know who your customer is. What is it that you actually want your system to be able to do in the real-world? After you have specified your target, you can tell me about the mechanics, the training procedures, and the state-of-the-art techniques. But first, we need a target worth aiming for.

Questions that a curious reader might have, which are not covered in this post:

- Why does Eliezer believe this?
- Is it never useful to have a better understanding of the mechanics, even if we don't have a clear target in mind?
- Do the mechanics of the system always depend on its target? Or are there some improvements in mechanics that could be robustly good across many (or all) targets?

*After the conversation, Eliezer mentioned that he often finds himself repeating this when he hears about alignment ideas. I asked him if he had written this up. He said no, but maybe someone like me should write it up.*

# Confused why a "capabilities research is good for alignment progress" position isn't discussed more

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

The predominant view on LW seems to be "pure AI capabilities research is bad, because capabilities progress alone doesn't contribute to alignment progress, and capabilities progress without alignment progress means that we're doomed".

I understand the arguments for this position, but I have what might be called the opposite position. The opposite position seems at least as intuitive as the standard position to me, and it confuses me that it's not discussed more. (I'm not confused that people reject it; I'm confused that nobody seems to even bring it up for the purpose of rejecting it.)

The opposite position is "In order to do alignment research, we need to understand how AGI works; and we currently don't understand how AGI works, so we need to have more capabilities research so that we would have a chance of figuring it out. Doing capabilities research now is good because it's likely to be slower now than it might be in some future where we had even more computing power, neuroscience understanding, etc. than we do now. If we successfully delayed capabilities research until a later time, then we might get a sudden spurt of it and wouldn't have the time to turn our increased capabilities understanding into alignment progress. Thus by doing capabilities research now, we buy ourselves a longer time period in which it's possible to do more effective alignment research."

Some reasons I have for holding this position:

**1)** I used to do AI strategy research. Among other things, I looked into [how feasible it is for intelligence to rapidly turn superintelligent](#), and [what kinds of pathways there are into AI disaster](#). But a thought that I kept having when doing any such research was "I don't know if any of this theory is of any use, because so much depends on what the world will be like when actual AGI is developed, and what that AGI will look like in the first place. Without knowing what AGI will look like, I don't know whether any of the assumptions I'm making about it are going to hold. If any one of them fails to hold, the whole paper might turn out to be meaningless."

Eventually, I concluded that I can't figure out a way to make the outputs of strategy research useful for as long as I know as little about AGI as I do. Then I went to do something else with my life, since it seemed too early to do useful AGI strategy research (as far as I could tell).

**2)** Compare the state of AI now, to how it was before the deep learning revolution happened. It seems obvious to me that our current understanding of DL puts us in a better position to do alignment research than we were before the DL revolution. For instance, Redwood Research is doing research on language models because they [believe that their research is analogous to some long-term problems](#).

Assume that Redwood Research's work will actually turn out to be useful for aligning superintelligent AI. Language models are one of the results of the DL revolution, so their work couldn't have been done before that revolution. It seems that in a counterfactual world where the DL revolution happened later and the DL era was compressed into a shorter timespan, our chances of alignment would be worse since that world's equivalent of Redwood Research would have less time to do their research.

**3)** As a similar consideration, language models are already "deceptive" in a sense - asked something that it has no clue about, InstructGPT will [happily come up with confident-sounding nonsense](#). When I linked people to some of that nonsense, multiple people pointed out that InstructGPT's answers sound like the kind of a student who's taking an exam and is asked to write an essay about a topic they know nothing about, but tries to fake it anyway (that is, trying to deceive the examiner).

Thus, even if you are doing pure capabilities research and just want your AI system to deliver people accurate answers, it is *already* the case that you can see a system like InstructGPT "trying to deceive" people. If you are building a question-answering system, you want to build one that people can trust to give accurate answers rather than impressive-sounding bullshit, so you have the incentive to work on identifying and stopping such "deceptive" computations as a capabilities researcher already.

So it has already happened that

- Progress in capabilities research gives us a new concrete example of how e.g. deception manifests in practice, that can be used to develop our understanding of it and develop new ideas for dealing with it.
- Capabilities research reaches a point where even capabilities researchers have a natural reason to care about alignment, reducing the difference between "capabilities research" and "alignment research".
- Thus, our understanding and awareness of deception is likely to improve as we get closer to AGI, and by that time we will have already learned a lot about how deception manifests in simpler systems and how to deal with it, and maybe some of that will suggest principles that generalize to more powerful systems as well.

It's not that I'd put a particularly high probability on InstructGPT by itself leading to any important insights about either deception in particular or alignment in general. InstructGPT is just an instance of something that seems likely to help us understand deception a little bit better. And given that, it seems reasonable to expect that further capabilities development will also give us small insights to various alignment-related questions, and maybe all those small insights will combine to give us the answers we need.

**4)** Still on the topic of deception, there are arguments suggesting that something like GPT will always be "deceptive" for [Goodhart's Law](#) and [Siren World](#) reasons. We can only reward an AI system for producing answers that look good to us, but this incentivizes the system to produce answers that look increasingly good to us, rather than answers that are actually correct. "Looking good" and "being correct" correlate with each other to some extent, but will eventually be pushed apart once there's enough optimization pressure on the "looking good" part.

As such, this seems like an unsolvable problem... *but* at the same time, if you ask me a question, I can have a desire to actually give a correct and useful answer to your

question, rather than just giving you an answer that you find maximally compelling. More generally, humans can and *often do* have a genuine desire to help other humans (or even non-human animals) fulfill their preferences, rather than just having a desire to superficially fake cooperativeness.

I'm not sure how this desire works, but I don't think you could train GPT to have it. It looks like some sort of theory of mind is involved in how the goal is defined. If I want to help you fulfill your preferences, then I have a sense of what it would mean for your preferences to be fulfilled, and I can have a goal of optimizing for that (even while I am uncertain of what exactly your preferences are).

We don't currently seem to know how to do this kind of a theory of mind, but it can't be *that* much more complicated than other human-level capabilities are, since even many non-human animals seem to have some version of it. Still, I don't think we can yet implement that kind of a theory of mind in any AI system. So we have to wait for our capabilities to progress to the kind of a point where this kind of a capacity becomes possible, and then we can hopefully use that capabilities understanding to solve what looks like a crucial piece of alignment understanding.

# Contra EY: Can AGI destroy us without trial & error?

NB: I've never published on LW before and apologize if my writing skills are not in line with LW's usual style. This post is an edited copy of the [same article](#) in my blog.

EY published an article last week titled "[AGI Ruin: A List of Lethalities](#)", which explains in detail why you can't train an AGI that won't try to kill you at the first chance it gets, as well as why this AGI will eventually appear given humanity's current trajectory in computer science. EY doesn't explicitly state a timeline over which AGI is supposed to destroy humanity, but it's implied that this will happen rapidly and humanity won't have enough time to stop it. EY doesn't find the question of *how exactly* AGI will destroy humanity too interesting and explains it as follows:

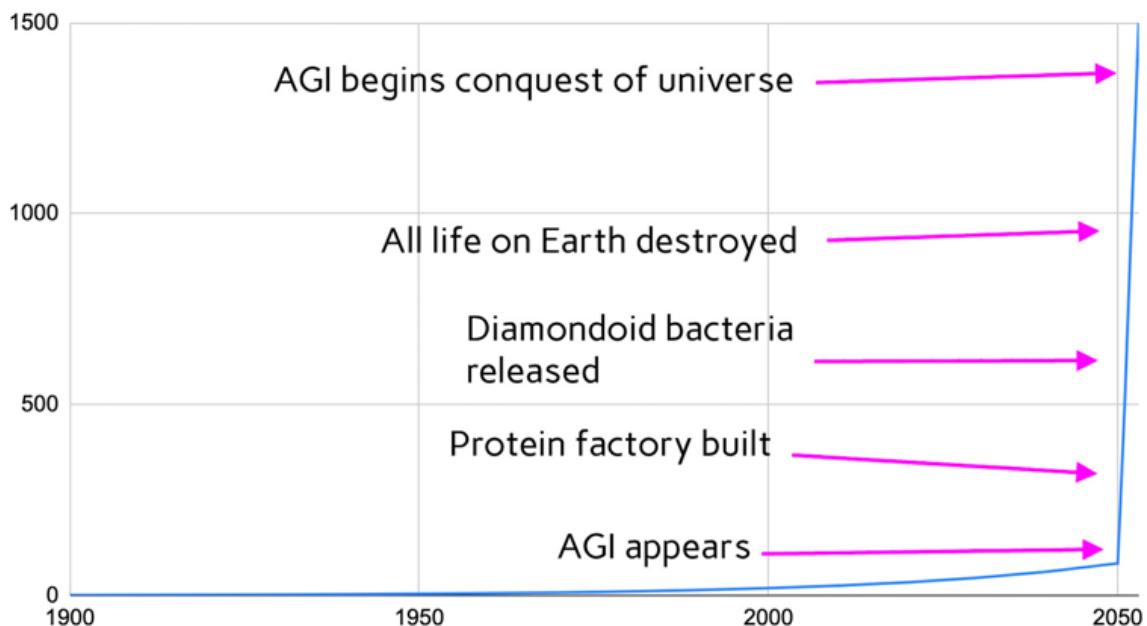
My lower-bound model of "how a sufficiently powerful intelligence would kill everyone, if it didn't want to not do that" is that it gets access to the Internet, emails some DNA sequences to any of the many many online firms that will take a DNA sequence in the email and ship you back proteins, and bribes/persuades some human who has no idea they're dealing with an AGI to mix proteins in a beaker, which then form a first-stage nanofactory which can build the actual nanomachinery. (Back when I was first deploying this visualization, the wise-sounding critics said "Ah, but how do you know even a superintelligence could solve the protein folding problem, if it didn't already have planet-sized supercomputers?" but one hears less of this after the advent of AlphaFold 2, for some odd reason.) The nanomachinery builds diamondoid bacteria, that replicate with solar power and atmospheric CHON, maybe aggregate into some miniature rockets or jets so they can ride the jetstream to spread across the Earth's atmosphere, get into human bloodstreams and hide, strike on a timer. **Losing a conflict with a high-powered cognitive system looks at least as deadly as "everybody on the face of the Earth suddenly falls over dead within the same second".**

Let's break down EY's proposed plan for "Skynet" into the requisite engineering steps:

1. Design a set of proteins that can form the basis of a "nanofactory"
2. Adapt the protein design to the available protein printers that accept somewhat-anonymous orders over the Internet
3. Design "diamondoid bacteria" that can kill all of humanity and that can be successfully built by the "nanofactory". The bacteria must be self replicating and able to extract power from solar energy for self sustainance.
4. Execute the evil plan by sending out the blueprints to unsuspecting protein printing corporations and rapidly taking over the world afterwards

The plan above looks great for a fiction book and EY is indeed a great [fiction writer](#) in addition to his Alignment work, but there's one unstated assumption: the AGI will not only be able to design everything using whatever human data it has available, but it will *also* execute the evil plan without needing lots of trial and error like mortal human inventors do. And surprisingly this part of EY's argument gets little objection. A visual representation of my understanding of EY's mental model of *AGI vs. Progress* is as follows:

Scientific progress vs. Year



## How fast can humans develop novel technologies?

Humans are the only known AGI that we have available for reference, so we could look at the fastest known examples of novel engineering to see how fast an AGI might develop something spectacular and human-destroying. Patrick Collison of Stripe keeps a helpful page titled “[Fast](#)” with notable “examples of people quickly accomplishing ambitious things together”. The engineering entries include:

- [P-80 Shooting Star](#), a World War II aircraft designed and built in 143 days by Lockheed Martin.
- [Spirit of St. Louis](#), another airplane designed and built in just 60 days.
- [USS Nautilus](#). The world’s first nuclear submarine was launched in 1173 days or 3.2 years.
- [Apollo 8](#), where it took 134 days between “what if we go to the moon?” to the actual landing.
- [The iPod](#), which took 290 days between the first designs and the device being launched to Apple stores.
- [Moderna’s vaccine](#) against COVID, which took 45 days between the virus being sequenced and the first batch of the actual vaccine getting manufactured.

Sounds very quick? Definitely, but the problem is that Patrick’s examples are all for engineering constructs building on top of decades of previous work. Designing a slightly better airplane in 1944 is not the same as creating the very first airplane in 1903, as by 1944 humans had 30 years of experience to build on top of. And if your task is to build *diamondoid bacteria manufactured by a protein-based nanomachinery factory* you’re definitely in Wright Brothers territory. So let’s instead look at timelines of *novel* technologies that had little prior research and infrastructure to fall back on:

- The Wright Brothers [took 4 years](#) to build their first successful prototype. It took [another 23 years](#) for the first mass manufactured airplane to appear, for a total of 27 years of R&D.
- It [took 63 years](#) for submarines to progress from “proof of concept” in the form of [Nautilus in 1800](#) to the first submarine capable of sinking a warship in the form of [The](#)

[Hunley](#) in 1863.

- It [took 40 years](#) between Einstein publishing his paper on the theory of relativity and the Atomic bomb being dropped on Hiroshima and Nagasaki. It took another 9 years to open the world's first [nuclear powerplant](#).
- It [took 36 years](#) from the first time mRNA vaccines were synthesized in 1984 and the first mRNA-based vaccine to be mass-manufactured.
- It [took at least 30 years](#) of development for LED technology to go from an experimental to being useful for commercial lighting.
- It took [around 30 years](#) for digital photography to overtake film photography in terms of costs and quality.

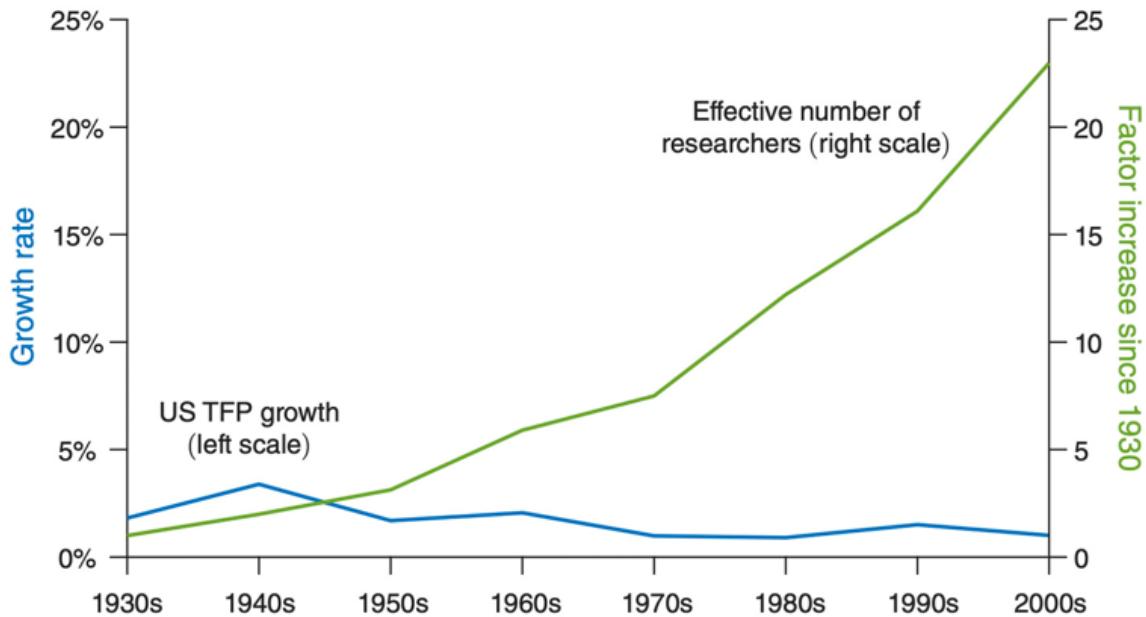
Now... you might object to this by correctly calling out the downside of human R&D:

- Human intellect is extremely inferior to what AGI will be capable of. At best, the collective intellectual capacity of the entire mankind will be equal to that of AGI. At worst, even all of our 8 billion brains will be collectively an order of magnitude dumber
- Humans have to sleep, eat, drink, vacation, while AGI can work 24/7
- Humans are more-or-less single threaded and require a coordinated effort to work on complicated research, which is additionally bogged down by the inefficiencies of trying to coordinate a large number of humans at the same time.

And this is all true! Humans are *nothing* to a hypothetical team of AGIs. But the problem is... until AGI can build its fantastical diamondoid bacteria, it remains dependent on imperfect human hands to conduct its R&D in the real world, as they'll be the only way for AGI to interact with the physical world for a very long time. Remember that AGI's one downside is that it will be running on motionless computers, unlike humans who have been running around with 4 limbs since the beginning of civilization. Which in turn brings us to the 30+ years timeline of developing a novel engineering construct, no matter how smart the AI will be.

## **Unstoppable intellect meets complexity of the universe**

Plenty of content has been written about how human scientific progress is slowing down, my favorite being [WTF Happened in 1971](#) and Scott's 2018 post [Is Science Slowing Down?](#). In the second article Scott brings up the paper [Are Ideas Getting Harder to Find?](#) by Bloom, Jones, Reenen & Webb (2018), which has the following neat graph:



We can see how the amount of investment into R&D is growing every year, but productive research is more or less flat. The paper brings up a relatable example in the section on semiconductor research:

The striking fact, shown in Figure 4, is that research effort has risen by a factor of 18 since 1971. This increase occurs while the growth rate of chip density is more or less stable: the constant exponential growth implied by Moore's Law has been achieved only by a massive increase in the amount of resources devoted to pushing the frontier forward. Assuming a constant growth rate for Moore's Law, the implication is that research productivity has fallen by this same factor of 18, an average rate of 6.8 percent per year. If the null hypothesis of constant research productivity were correct, the growth rate underlying Moore's Law should have increased by a factor of 18 as well. Instead, it was remarkably stable. Put differently, because of declining research productivity, **it is around 18 times harder today to generate the exponential growth behind Moore's Law than it was in 1971.**

Not even AGI could get around this problem and would likely require an exponentially growing amount of resources as it delves deeper into engineering and fundamental research. It is definitely true that AGI *itself* will be rapidly increasing its intellect, but can this really continue indefinitely? At some point all the low hanging fruit missed by human AI researchers will be exhausted and AGI will have to spend years in real world time to make significant improvements of its own IQ. Granted, AGI *will* rapidly reach an IQ far beyond human reach, but all this intellectual power will still have to contend with the difficulties of novel research.

## What does AGI want?

Since AGI development is completely decoupled from mammalian evolution here on Earth, its quite likely to eventually exhibit "[blue and orange](#)" morality, behaving in a completely alien and unpredictable fashion, with no humanly understandable motivations or a way for humans to relate to what the AGI wants. That being said, AGI is likely to fall into one of two buckets regardless of its motivations:

1. AGI will act rationally to achieve whatever internal goals it has, no matter how alien and weird to us. I.e. “collect all the shiny objects into every bucket-like object in the universe” or “[convert the universe into paperclips](#)”. This means the AGI will carefully plan ahead and attempt to preserve its own existence to fulfill the internal overreaching goals.
2. AGI doesn’t have any goals at all beyond “kill all humans!”. It just acts as a rogue terrorist, attempting to destroy humans without the slightest concern for its own survival. If all humans die and the AGI dies alongside them, that’s fine according to the AGI’s internal motivations. There’s no attempt to ensure overarching continuation of its goals (like “[collect all strawberries!](#)”) once humanity is extinct.

Let’s start with scenario #1 by looking at... the common pencil.

## What does it take to make a pencil?

A classic pamphlet called [I, Pencil](#) walks us through what it takes to make a common pencil from scratch:

1. Trees have to be cut down, which takes saws, trucks, rope and countless other gear.
2. The cut down trees have to be transported to the factory by rail, which in turns needs laid down rail, trains, rail stations, cranes, etc.
3. The trees are cut down with metal saws, waxes and dried. This consumes a lot of electricity, which is in turn made by burning fuel, making solar panel or building hydroelectric powerplants.
4. At the center of the pencil is a chunk of graphite mined in Sri Lanka, using loads of equipment and transported by ship.
5. The pencil is covered with lacquer, that’s in turn made by growing castor beans.
6. There’s the piece of metal holding the eraser, mined from shafts underground.
7. And finally there’s the rubber mined in Indonesia and once again transported by ship.

The point to this entire story is that making something as simple as a pencil requires a **massive** supply chain employing tens of millions of non-AGI humans. If you want any hope of continuing to exist, you need to replace the labor of this gigantic global army of humans with AGI-controlled robots or “diamondoid bacteria” or whatever other magical contraption you want to invoke. Which will require lots of trial & error and decades of building out a reliable AGI-controlled supply chain that could be reused to fight humans at the drop of a hat. Because otherwise AGI will risk seeing its brilliant plan fail, resulting in humans going berserk against any machines capable of running said AGI and ending its reign of Earth long before it has a chance to start in earnest. And if the AGI doesn’t understand this... how smart is it really?

## YOLO AGI?

But what if the AGI is absolutely ruthless and doesn’t care if it goes up in flames as soon as humans are gone? Then we could get to the end of humanity much faster with options like:

- Get humans to think that their enemy is about to launch a nuclear strike and launch a strike of their own, similar to [WarGames](#)
- Design a supervirus capable of destroying humanity. Think a combination of HIV’s lethality with the ease of spread of measles.
- Plant a powerful [information hazard](#) into humanity’s consciousness that will somehow trigger us to kill each other as soon as possible. Also see [Roko’s Basilisk](#) and [Rokoko’s Basilisk](#), an infohazard responsible for the birth of [X AE A-12](#).
- Design the mother of all [greenhouse gases](#) and convince humanity to make tons of it, eventually resulting in the planet heating up to extreme temperatures.
- Provide advanced nuke designs and materials covertly to very bad people and manipulate them into sabotaging world order.

The problem with all these scenarios is similar:

Either they're perfectly doable by humans in the present, with no AGI help necessary. I.e. we've been [barely saved](#) from WW3 by a Soviet officer, long before AGI was on anyone's mind. So at worst AGI will somewhat increase the risks of this happening in the short term... Or they require lots of trial & error to develop into functional production-ready technologies, once again creating a big problem for AGI, as it has to rely on imperfect humans to do the novel R&D. This will still take decades, even if AGI won't worry about a full takeover of supply chains.

## But what about AlphaFold?

Another possible counter-argument is that AGI will figure out the laws of the universe through internal modeling and will be able to simulate and perfect its amazing inventions without needing trial & error in the physical world. EY mentions [AlphaFold](#) as an example of such a breakthrough. If you haven't heard about it, here's a description of the Protein Folding Problem from Wiki that AlphaFold 2 solved better than any other prior system back in 2020:

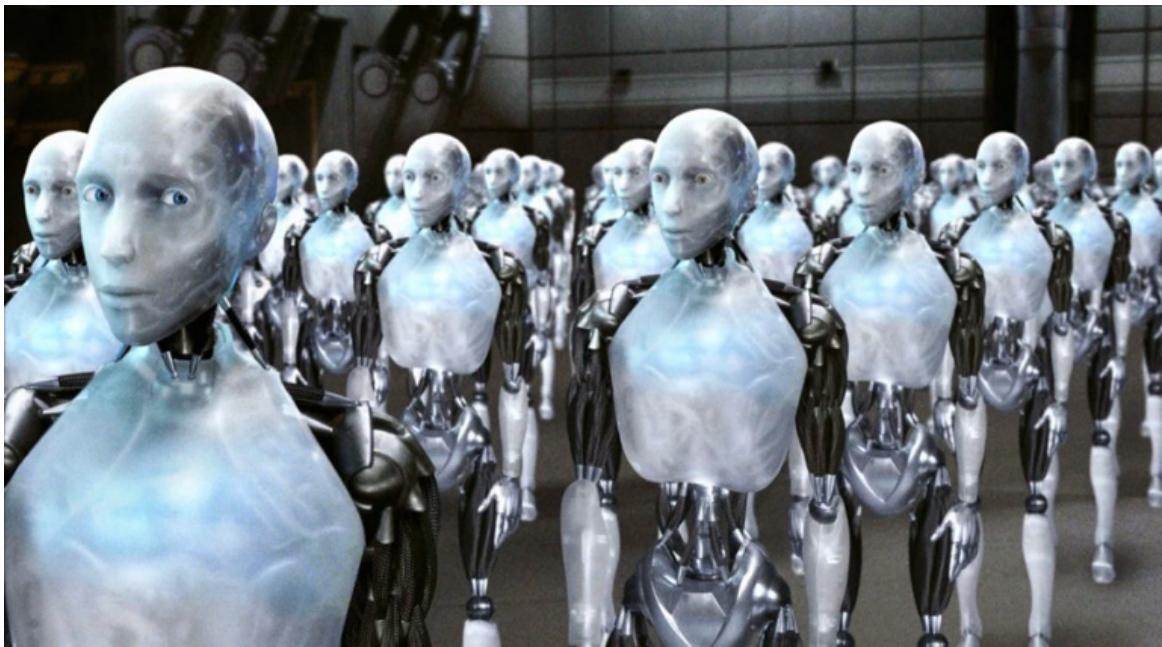
Proteins consist of chains of amino acids which spontaneously fold, in a process called protein folding, to form the three dimensional (3-D) structures of the proteins. The 3-D structure is crucial to the biological function of the protein. However, understanding how the amino acid sequence can determine the 3-D structure is highly challenging, and this is called the "protein folding problem". The "protein folding problem" involves understanding the thermodynamics of the interatomic forces that determine the folded stable structure, the mechanism and pathway through which a protein can reach its final folded state with extreme rapidity, and how the native structure of a protein can be predicted from its amino acid sequence.

According to EY, the existence of AlphaFold shows that a smart enough AGI could eventually learn to manipulate proteins into "nanofactories" that could be used to interact with the physical world. However the current version still [has major limitations](#):

Whilst it may be considered the gold standard of protein prediction, there is still room for improvement as AlphaFold only provides one prediction of a stable structure for each protein; however, proteins are dynamic and can change shape throughout the body, for example under different pH conditions. Additionally, AlphaFold is not able to determine the shape of multiprotein complexes and does not include any ligands such as cofactors or metals, meaning no data are available for such interactions. Despite these shortcomings, AlphaFold is the first step in protein prediction technology, and it is likely that solving these challenges will also be done so using deep learning and AI.

In other words, there's still a **huge** leap between "can predict simple protein structures" and "can design protein nanofactories without experimentation". AGI will likely need to spend *decades* managing laboratory experiments to fill in the gaps around our understanding of how proteins work. And don't forget that currently available commercial protein printers are not perfect, especially if you're trying to print a novel structure of far bigger complexity than anything else on the planet. Also see this [excellent comment](#) on the subject by [anonymousaisafety](#).

## What if AGI settles for a robot army?



*Cybernetic army from I, Robot*

We could also think of the *diamondoid bacteria* as just an example of what the AI can do and turn to other ways it could manipulate the physical reality, that are closer to the technology that we already have today. There's [impressive videos](#) of Boston Dynamics robots doing all kinds of stunts, so we could ask if perhaps AGI could utilize their existing progress to quickly give itself a way to interact with the outside world. However this would still involve many roadblocks:

- The robots are running on pre-programmed runs in a human-designed course and are not capable of navigating through unknown terrain. AGI will have to push cybernetics to amazing heights in order for the robots to be useful in novel spaces.
- AGI *could* just make a version of itself and give sentience to every individual machine, but this will require a massive leap in computational technology, as a single machine is very unlikely to be able to host an instance of an AGI.
- AGI *could* have a command and control center next to every batch of robots or try to control them over large distances, but this is again a daunting engineering task when you have little room for mistake.
- Battery technology is currently too lousy and even the more simple dog-like robots can only last for [90 minutes](#). There's no published information on how long the human-like versions could last on battery, but we could reason it cannot be more than an hour. Therefore AGI will need to achieve a big leap in battery life before anything resembling the "I, Robot" machinery could be built.
- Boston Dynamics robots don't exist in large enough quantities for a complete overtake of global supply chains and it will take years of *human labor* to get more robot factories online.
- Humans are pretty good at warfare and your robots have to be *extremely* good to beat them in battle, far better than what's currently available.

[added] Also see this [excellent comment](#) by *anonymousaisafety* explaining why "just takeover the human factories" is not a quick path to success (slightly edited below):

The tooling and structures that a superintelligent AGI would need to act autonomously does not actually exist in our current world, so before we can be made into paperclips, there is a necessary period of bootstrapping where the superintelligent AGI designs and manufactures new machinery using our current machinery. Whether it's an unsafe AGI

that is trying to go rogue, or an aligned AGI that is trying to execute a "pivotal act", the same bootstrapping must occur first.

Case study: a common idea I've seen while lurking on LessWrong and SSC/ACT for the past N years is that an AGI will "just" hack a factory and get it to produce whatever designs it wants. This is not how factories work. There is no 100% autonomous factory on Earth that an AGI could just take over to make some other widget instead. Even [highly automated](#) factories are:

1. Highly automated to produce a specific set of widgets,
2. Require physical adjustments to make different widgets, and...
3. Rely on humans for things like input of raw materials, transferring in-work products between automated lines, and the testing or final assembly of completed products. 3D printers are one of the worst offenders in this regard. The public perception is that a 3D printer can produce anything and everything, but they actually have pretty strong constraints on what types of shapes they can make and what materials they can use, and usually require multi-step processes to avoid those constraints, or post-processing to clean up residual pieces that aren't intended to be part of the final design, and almost always a 3D printer is producing sub-parts of a larger design that still must be assembled together with bolts or screws or welds or some other fasteners.

So if an AGI wants to have unilateral control where it can do whatever it wants, the very first prerequisite is that it needs to create a futuristic, fully automated, fully configurable, network-controlled factory -- which needs to be built with what we have now, and that's where you'll hit the supply constraints for things like lead times on part acquisition. The only way to reduce this bootstrapping time is to have this stuff designed in advance of the AGI, but that's backwards from how modern product development actually works. We design products, and then we design the automated tooling to build those products. If you asked me to design a factory that would be immediately usable by a future AGI, I wouldn't know where to even start with that request. I need the AGI to tell me what it wants, and then I can build that, and then the AGI can takeover and do their own thing.

A related point that I think gets missed is that our automated factories aren't necessarily "fast" in a way you'd expect. There's long lead times for complex products. If you have the specialized machinery for creating new chips, you're still looking at ~14-24 weeks from when raw materials are introduced to when the final products roll off the line. We hide that delay by constantly building the same things all of the time, but it's very visibly when there's a sudden demand spike -- that's why it takes so long before the supply can match the demand for products like processors or GPUs. I have no trouble with imagining a superintelligent entity that could optimize this and knock down the cycle time, but there's going to be physical limits to these processes and the question is can it knock it down to 10 weeks or to 1 week? And when I'm talking about optimization, this isn't just uploading new software because that isn't how these machines work. It's designing new, faster machines or redesigning the assembly line and replacing the existing machines, so there's a minimum time required for that too before you can benefit from the faster cycle time on actually making things. Once you hit practical limits on cycle time, the only way to get more stuff faster is to scale wide by building more factories or making your current factories even larger.

If we want to try and avoid the above problems by suggesting that the AGI doesn't actually hack existing factories, but instead it convinces the factory owners to build the things it wants instead, there's not a huge difference -- instead of the prerequisite here being "build your own factory", it's "hostile takeover of existing factory", where that hostile takeover is either done by manipulation, on the public market, as a private sale, or by outbidding existing customers (e.g. have enough money to convince TSMC to make your stuff instead of Apple's), or with actual arms and violence. There's still the other

lead times I've mentioned for retooling assembly lines and actually building a complete, physical system from one or more automated lines.

My prediction is that it will take AGI *at least* 30 years of effort to get to a point where it can comfortably rely on the robots to interact with the physical world and not have to count on humans for its supply chain needs.

## [added] What if AGI just simulates our physical world?

This idea goes hand-in-hand with idea that AlphaFold is the answer to all challenges in bioengineering. There are two separate assumptions here, both found in the field of computational complexity:

1. That an AGI can simulate the physical systems *perfectly*, i.e. physical systems are computable processes.
2. That an AGI can simulate the physical systems *efficiently*, i.e. either  $P = NP$ , or for some reason all of these interesting problems that the AGI is solving are NOT known to be isomorphic to some NP-hard problem.

I don't think these assumptions are reasonable. For a full explanation see [this excellent comment](#) by anonymousaisafety.

## Mere mortals can't comprehend AGI?

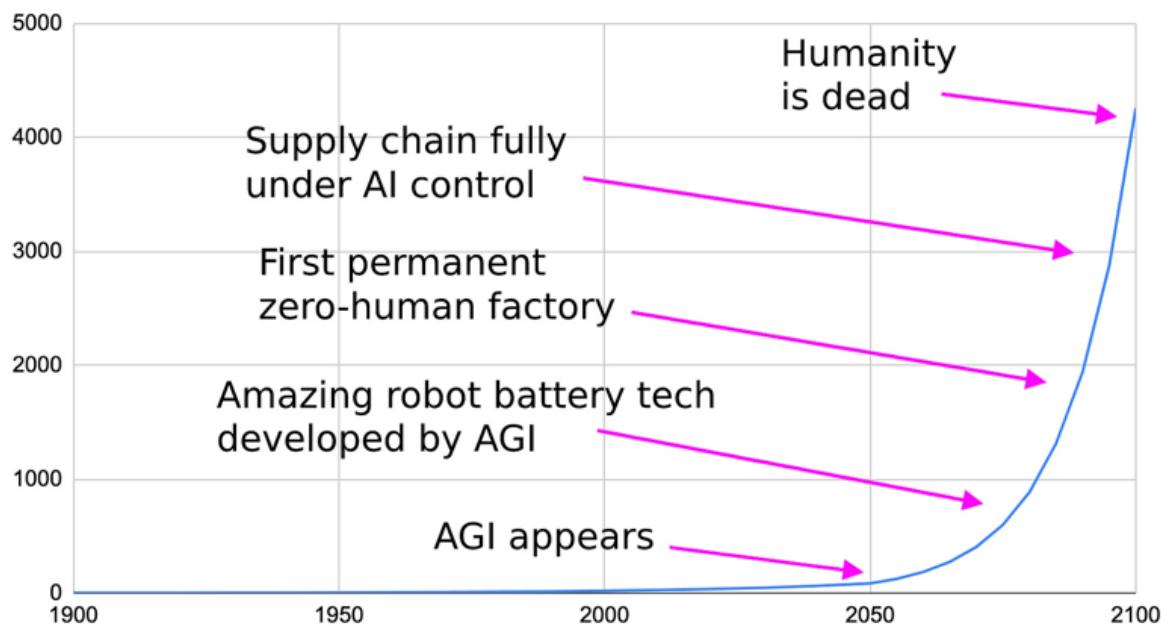
Another argument is that AGI will achieve such an incomprehensible level of intellect that it will become impossible to predict what it will be capable of. I mean, who knows, maybe with an IQ of 500 you could just magically turn yourself into a God and destroy Earth with a [Thanos-style snap](#) of your fingers? But I contend that even a creature with an IQ of 500 will be inherently limited by our physical universe and won't magically become gain [omniscience](#) by virtue of its intellect alone. It will instead have to spend decades to get rid of using humans as a proxy, no matter how smart it could be potentially.

## Does this mean EY is wrong and AGI is not a threat?

I believe that EY is only wrong about handwaving the difficulties of growing from a computer-based AGI to an AGI capable of operating independently from the human race. In the long-term his predictions will likely come true, once AGI has enough time to go through the difficult R&D cycle of building the nanofactories and diamondoid bacteria. My predicted timeline is as follows:

1. AGI first appears somewhere around 2040, in line with the [Metaculus prediction](#).
2. After a few years of peaceful coexistence, AI Alignment researchers are mocked for their doomer predictions and everyone thinks that AGI is perfectly safe. EY will keep writing blog posts about how everyone is wrong and AGI cannot be trusted. AGI might start working behind the shadows to try and get AI Alignment researchers silenced.
3. AGI spends decades convincing humanity to let it take over the global supply chains and to run complex experiments to manufacture advanced AGI-designed machinery, supposedly necessary to improve human living standards. This will likely take *at least* 30 years, as per our reference to how long it took to implement other gigantic breakthroughs in science.
4. Once the AGI is convinced that all the cards have fallen into place and humans could be safely removed, it will pull the plug and destroy us all.

Scientific progress vs. Year



*Updated version of the original progress graph*

I'm hoping that the AI Alignment movement tries to spend more time on the low level engineering details of "humanity goes poof" rather than handwaving everything away via science fiction concepts. Because otherwise it's hard to believe that the FOOM scenario could ever come to fruition. And if FOOM is not the real problem, perhaps we could save humanity by managing AGI's interactions with the physical world more carefully once it appears?

# Intergenerational trauma impeding cooperative existential safety efforts

**Epistemic status:** personal judgements based on conversations with ~100 people aged 30+ who were worried about AI risk "before it was cool", and observing their effects on a generation of worried youth, at a variety of EA-adjacent and rationality-community-adjacent events.

**Summary:** There appears to be something like inter-generational trauma among people who think about AI x-risk — including some of the AI-focussed parts of the EA and rationality communities — which is

- preventing the formation of valuable high-trust relationships with newcomers that could otherwise be helpful to humanity collectively making better decisions about AI, and
- feeding the formation of small pockets of people with a highly adversarial stance towards the rest of the world (and each other).

[This post is also available on the [EA Forum](#).]

## **Part 1 — The trauma of being ignored**

You — or some of your close friends or colleagues — may have had the experience of fearing AI would eventually pose an existential risk to humanity, and trying to raise this as a concern to mainstream intellectuals and institutions, but being ignored or even scoffed at just for raising it. That sucked. It was not silly to think AI could be a risk to humanity. It can.

I, and around 100 people I know, have had this experience.

Experiences like this can easily lead to an attitude like “Screw those mainstream institutions, they don’t know anything and I can’t trust them.”

At least 30 people I've known personally have adopted that attitude in a big way, and I estimate many more. In the remainder of this post, I'd like to point out some ways this attitude can turn out to be a mistake.

## **Part 2 — Forgetting that humanity changes**

Basically, as AI progresses, it becomes easier and easier to make the case that it could pose a risk to humanity's existence. When people didn't listen about AI risks in the past, that happened under certain circumstances, with certain AI capabilities at the forefront and certain public discourse surrounding them. These circumstances have changed, and will continue to change. It may not be getting easier as fast as one would ideally like, but it is getting easier. Like the stock market, it may be hard to predict how and when things will change, but they will.

If one forgets this, one can easily adopt a stance like "mainstream institutions will never care" or "the authorities are useless". I think these stances are often exaggerations of the truth, and if one adopts them, one loses out on the opportunity to engage productively with the rest of humanity as things change.

## **Part 3 - Reflections on the Fundamental Attribution Error (FAE)**

The Fundamental Attribution Error ([wiki/Fundamental\\_attribution\\_error](#)) is a cognitive bias whereby you too often attribute someone else's behavior to a fundamental (unchanging) aspect of their personality, rather than considering how their behavior might be circumstantial and likely to change. With a moment's reflection, one can see how the FAE can lead to

- trusting too much — assuming someone would never act against your interests because they didn't the first few times, and also
- trusting too little — assuming someone will never do anything good for you because they were harmful in the past.

The second reaction could be useful for getting out of abusive relationships. The risk of being mistreated over and over by someone is usually not worth the opportunity cost of finding new people to interact with. So, in personal relationships, it can be healthy to just think "screw this" and move on from someone when they don't make a good first (or tenth) impression.

## **Part 4 — The FAE applied to humanity**

If one has had the experience of being dismissed or ignored for expressing a bunch of reasonable arguments about AI risk, it would be easy to assume that humanity (collectively) can never be trusted to take such arguments seriously. But,

1. Humanity has changed greatly over the course of history, arguably more than any individual has changed, so it's suspect to assume that humanity, collectively, can never be rallied to take a reasonable action about AI.
2. One does not have the opportunity to move on and find a different humanity to relate to. "Screw this humanity who ignores me, I'll just imagine a different humanity and relate to that one instead" is not an effective strategy for dealing with the world.

## **Part 5 - What, if anything, to do about this**

If the above didn't resonate with you, now might be a good place to stop reading :)  
Maybe this post isn't good advice for you to consider after all.

But if it did resonate, and you're wondering what you may be able to do differently as a result, here are some ideas:

- Try saying something nice and civilized about AI risk that you used to say 5-10 years ago, but which wasn't well received. Don't escalate it to something more offensive or aggressive; just try saying the same thing again. Someone new might take interest today, who didn't care before. This is progress. This is a sign that humanity is changing, and adapting somewhat to the circumstances presented by AI development.
- Try Googling a few AI-related topics that no one talked about 5-10 years ago to see if today more people are talking about one or more of those topics. Switch up the keywords for synonyms. (Maybe keep a list of search terms you tried so you don't go in circles, and if you really find nothing, you can share the list and write an interesting LessWrong post speculating about why there are no results for it.)
- Ask yourself if you or your friends feel betrayed by the world ignoring your concerns about AI. See if you have a "screw them" feeling about it, and if that feeling might be motivating some of your discussions about AI.
- If someone older tells you "There is nothing you can do to address AI risk, just give up", maybe don't give up. Try to understand their experiences, and ask yourself seriously if those experiences could turn out differently for you.

# why assume AGIs will optimize for fixed goals?

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

When I read posts about AI alignment on LW / AF/ Arbital, I almost always find a particular bundle of assumptions taken for granted:

- An AGI has a single terminal **goal**<sup>[1]</sup>.
- The goal is a **fixed** part of the AI's structure. The internal dynamics of the AI, if left to their own devices, will never modify the goal.
- The "outermost loop" of the AI's internal dynamics is an optimization process aimed at the goal, or at least the AI behaves just as though this were true.
- This "outermost loop" or "**fixed-terminal-goal-directed wrapper**" chooses which of the AI's specific capabilities to deploy at any given time, and how to deploy it<sup>[2]</sup>.
- The AI's capabilities will themselves involve optimization for sub-goals that are not the same as the goal, and they will optimize for them very powerfully (hence "capabilities"). **But it is "not enough" that the AI merely be good at optimization-for-subgoals: it will also have a fixed-terminal-goal-directed wrapper.**
  - So, the AI may be very good at playing chess, and when it is playing chess, it may be running an internal routine that optimizes for winning chess. This routine, and not the terminal-goal-directed wrapper around it, explains the AI's strong chess performance. ("Maximize paperclips" does not tell you how to win at chess.)
  - The AI may also be good at things that are much more general than chess, such as "planning," "devising proofs in arbitrary formal systems," "inferring human mental states," or "coming up with parsimonious hypotheses to explain observations." All of these are capacities<sup>[3]</sup> to optimize for a particular subgoal that is not the AI's terminal goal.
  - Although these subgoal-directed capabilities, and not the fixed-terminal-goal-directed wrapper, will constitute the *reason* the AI does well at anything it does well at, the AI must **still** have a fixed-terminal-goal-directed wrapper around them and apart from them.
- There is no way for the terminal goal to change through bottom-up feedback from anything inside the wrapper. The hierarchy of control is strict and only goes one way.

My question: why assume all this? Most pressingly, why assume that the terminal goal is fixed, with no internal dynamics capable of updating it?

I often see the rapid capability gains of humans over other apes cited as a prototype case for the rapid capability gains we expect in AGI. But humans do *not* have this wrapper structure! Our goals often change over time. (And we often permit or even welcome this, whereas an optimizing wrapper would try to prevent its goal from changing.)

Having the wrapper structure was evidently not necessary for *our* rapid capability gains. Nor do I see reason to think that our capabilities result from us being "more

structured like this" than other apes. (Or to think that we are "more structured like this" than other apes in this first place.)

Our capabilities seem more like the subgoal capabilities discussed above: general and powerful tools, which can be "plugged in" to many different (sub)goals, and which do not require the piloting of a wrapper with a fixed goal to "work" properly.

Why expect the "wrapper" structure with fixed goals to emerge from an outer optimization process? Are there any relevant examples of this happening via natural selection, or via gradient descent?

There are many, many posts on LW / AF/ Arbital about "optimization," its relation to intelligence, whether we should view AGIs as "optimizers" and in what senses, etc. I have not read all of it. Most of it touches only lightly, if at all, on my question. For example:

- There has been much discussion over whether an AGI would inevitably have (close to) consistent preferences, or would self-modify itself to have closer-to-consistent preferences. See e.g. [here](#), [here](#), [here](#), [here](#). Every post I've read on this topic implicitly assumes that the preferences are fixed in time.
- Mesa-optimizers have been discussed extensively. The same bundle of assumptions is made about mesa-optimizers.
- It has been [argued](#) that if you *already* have the fixed-terminal-goal-directed wrapper structure, then you will prefer to avoid outside influences that will modify your goal. This is true, but does not explain why the structure would emerge in the first place.
- There are arguments ([e.g.](#)) that we should *heuristically* imagine a superintelligence as a powerful optimizer, to get ourselves to predict that it will not do things we know are suboptimal. These arguments tell us to imagine the AGI picking actions that are optimal for a goal iff it is currently optimizing for that goal. They don't tell us when it will be optimizing for which goals.

---

EY's notion of "consequentialism" seems closely related to this set of assumptions. But, I can't extract an answer from the writing I've read on that topic.

EY [seems to attribute](#) what I've called the powerful "subgoal capabilities" of humans/AGI to a property called "cross-domain consequentialism":

We can see one of the critical aspects of human intelligence as [cross-domain consequentialism](#). Rather than only forecasting consequences within the boundaries of a narrow domain, we can trace chains of events that leap from one domain to another. Making a chess move wins a chess game that wins a chess tournament that wins prize money that can be used to rent a car that can drive to the supermarket to get milk. An Artificial General Intelligence that could learn many domains, and engage in consequentialist reasoning that leaped across those domains, would be a [sufficiently advanced agent](#) to be interesting from most perspectives on interestingness. It would start to be a consequentialist about the real world.

while defining "consequentialism" as the ability to do means-end reasoning with some preference ordering:

Whenever we reason that an agent which prefers outcome Y over Y' will therefore do X instead of X' we're implicitly assuming that the agent has the cognitive

ability to do consequentialism at least about Xs and Ys. It does means-end reasoning; it selects means on the basis of their predicted ends plus a preference over ends.

But the ability to use this kind of reasoning, and do so across domains, does not imply that one's "outermost loop" looks like this kind of reasoning applied to the whole world at once.

I myself am a cross-domain consequentialist -- a human -- with very general capacities to reason and plan that I deploy across many different facets of my life. But I'm not running an outermost loop with a fixed goal that pilots around all of my reasoning-and-planning activities. Why can't AGI be like me?

EDIT to spell out the reason I care about the answer: agents with the "wrapper structure" are inevitably hard to align, in ways that agents without it might not be. An AGI "like me" might be morally uncertain like I am, persuadable through dialogue like I am, etc.

It's very important to know what kind of AIs would or would not have the wrapper structure, because this makes the difference between "inevitable world-ending nightmare" and "we're not the dominant species anymore." The latter would be pretty bad for us too, but there's a difference!

1. ^

Often people speak of the AI's "utility function" or "preference ordering" rather than its "goal."

For my purposes here, these terms are more or less equivalent: it doesn't matter whether you think an AGI must have *consistent* preferences, only whether you think it must have *fixed* preferences.

2. ^

...or at least the AI behaves just as though this were true. I'll stop including this caveat after this.

3. ^

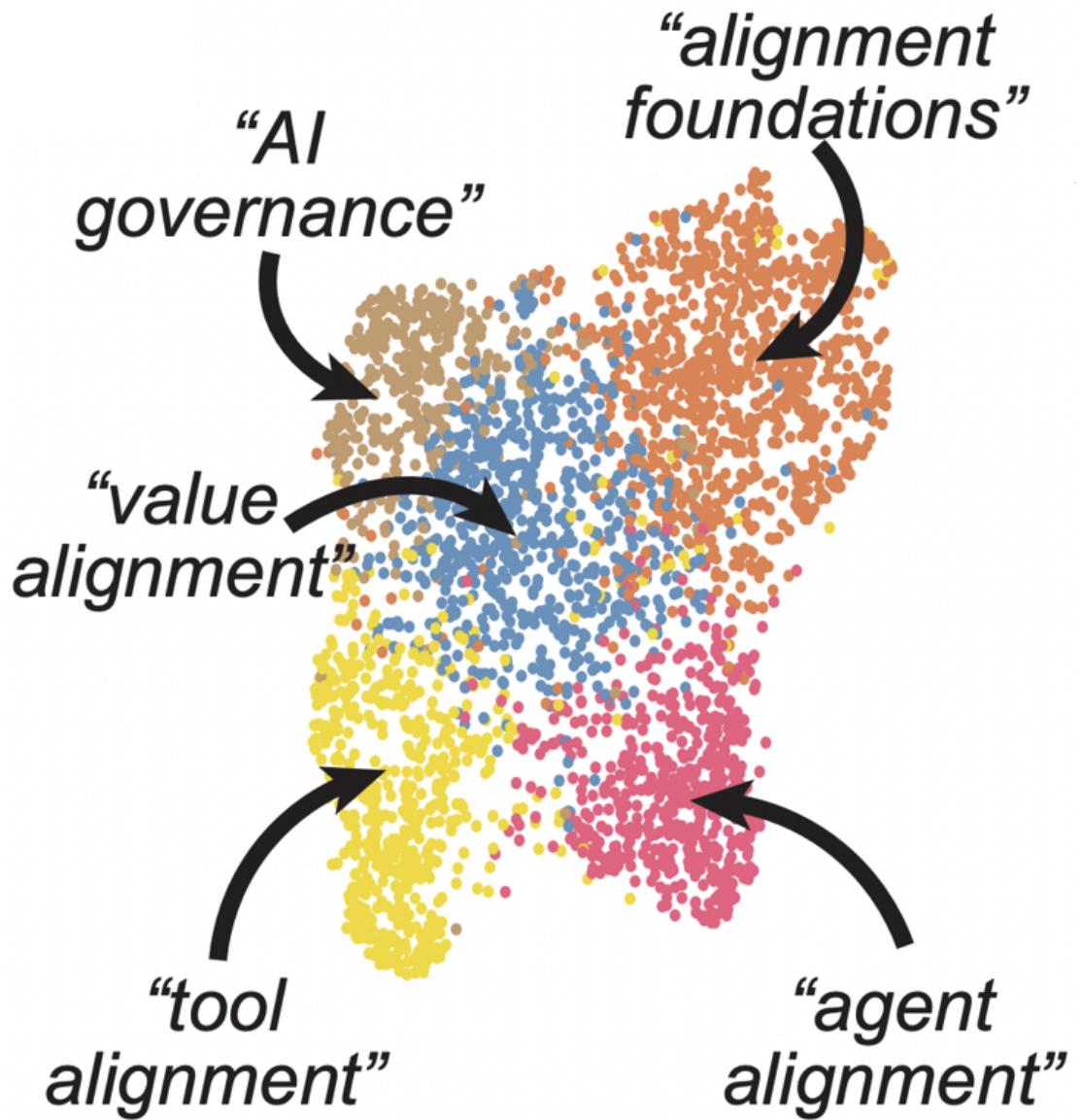
Or possibly one big capacity -- "general reasoning" or what have you -- which contains the others as special cases. I'm not taking a position on how modular the capabilities will be.

# A descriptive, not prescriptive, overview of current AI Alignment Research

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

*TL;DR: In this project, we collected and cataloged AI alignment research literature and analyzed the resulting dataset in an unbiased way to identify major research directions. We found that the field is growing quickly, with several subfields emerging in parallel. We looked at the subfields and identified the prominent researchers, recurring topics, and different modes of communication in each. Furthermore, we found that a classifier trained on AI alignment research articles can detect relevant articles that we did not originally include in the dataset.*

(video presentation [here](#))



## Dataset Announcement

In the context of the [6th AISc](#), we collected a dataset of alignment research articles from a variety of different sources. This dataset is now available for download [here](#) and the code for reproducing the scrape is on GitHub [here<sup>\[1\]</sup>](#). When using the dataset, please cite our manuscript as described in the footnote<sup>[2]</sup>.

source	domain	# of articles
Alignment Forum	alignmentforum.org	2,138
	lesswrong.com	28,252
arXiv	AI alignment research (level-0)	707
	AI research (level-1)	1,679
	arXiv.org/search/?query=quantum	1,000
	arXiv.org/list/cs.AI (filtered)	4,621
Books	(available upon request)	23
Blogs	aiimpacts.org	227
	aipulse.org	23
	aisafety.camp	8
	carado.moe	59
	cold-takes.com	111
	deepmindsafetyresearch.medium.com	10
	generative.ink	17
	gwern.net	7
	intelligence.org	479
	jsteinhardt.wordpress.com	39
	qualiacomputing.com	278
	vkrakovna.wordpress.com	43
	waitbutwhy.com	2
	yudkowsky.net	23
Newsletter	rohinshah.com/alignment-newsletter/ summaries	420
Reports	pdf-only articles	323
	distill.pub	49
Audio transcripts	youtube.com playlist 1 & 2	457
	Assorted transcripts	25
	interviews with AI researchers <sup>33</sup>	12
Wikis	arbital.com	223
	lesswrong.com (Concepts Portal)	227
	stampy.ai	132
Total:	Total token count: 89,240,129 Total word count: 53,550,146 Total character count: 351,757,163	

**Table 1: Different sources of text included in the dataset alongside the number of articles per source.** Color of row indicates that data was analyzed as AI alignment research articles (green) or baseline (gray), or that the articles were added to the dataset as a result of the analysis in Fig. 4 (purple). Definition of level-0 and level-1 articles in Fig. 4c. For details about our collection procedure see the Methods section.

Here follows an abbreviated version of the full [manuscript](#), which contains additional analysis and discussion.

# Rapid growth of AI Alignment research from 2012 to 2022 across two platforms

After collecting the dataset, we analyzed the two largest non-redundant sources of articles, Alignment Forum (AF) and arXiv. We found rapid growth in publications on the AF (Fig. 1a) and a long-tailed distribution of articles per researcher (Fig. 1b) and researchers per article (Fig. 1c). We were surprised to find a *decrease* in publications on the arXiv in recent years, but identified the cause for the decrease as spurious and fixed the issue in the published dataset (details in Fig. 4).

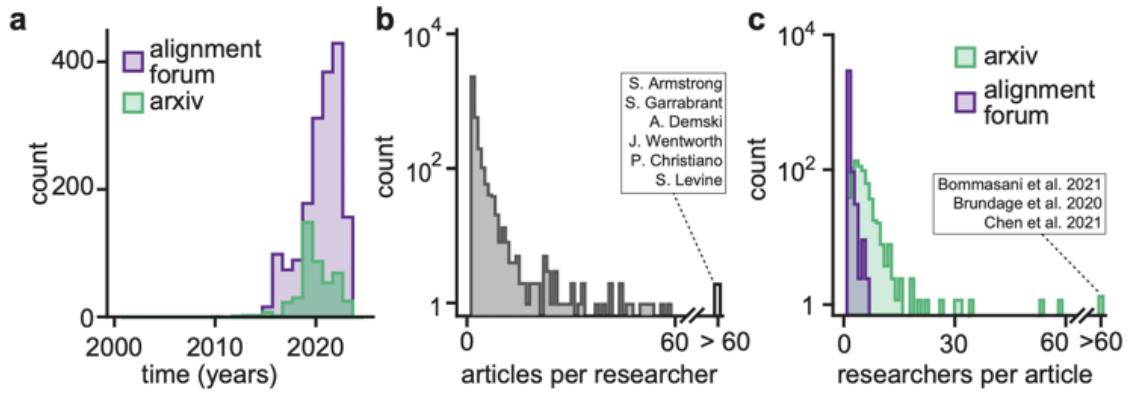
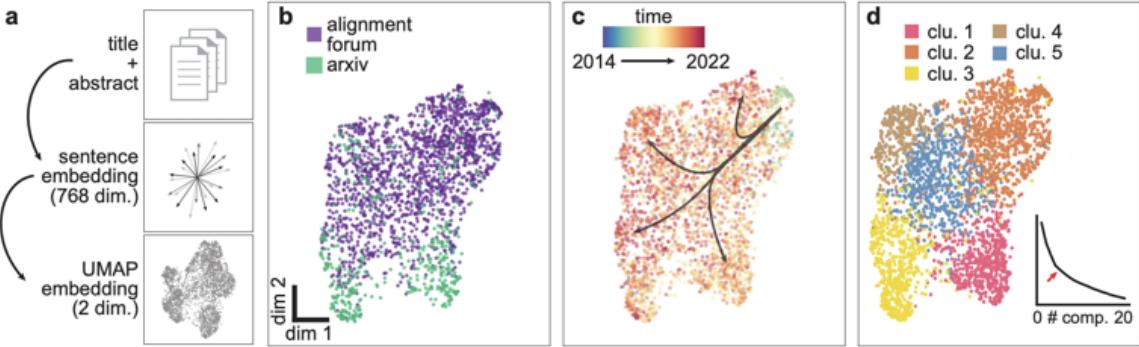


Figure 1: **Alignment Research across a community forum and a preprint server.** (a) Number of articles published as a function of time on the alignment forum (AF; purple) and the arXiv preprint server (arXiv; green). (b) Histogram of the number of articles per researcher published on either AF or arXiv. Inset shows names of six researchers with more than 60 articles. Note the logarithmic y-axis. (c) Histogram of the number of researchers per article on AF (purple) and arXiv (green). Note the logarithmic y-axis.

## Unsupervised decomposition of AI Alignment research into distinct clusters

Given access to this unique dataset, we were curious to see if we could identify distinct clusters of research. We mapped the title + abstract of each article into vector form using the [Allen Institute for AI's SPECTER model](#) and reduced the dimensionality of the embedding with UMAP (Fig. 2a). The resulting manifold shows a continuum of AF posts and arXiv articles (Fig. 2b) and a temporal gradient from the top right to the bottom left (Fig. 2c). Using k-means and the [elbow method](#), we obtain five clusters of research articles that map onto distinct regions of the UMAP projection (Fig. 2d).



**Figure 2: Dimensionality reduction and unsupervised clustering of alignment research.** (a) Schematic of the embedding and dimensionality reduction. After concatenating title and abstract of articles, we embed the resulting string with the Allen SPECTER model40, and then perform UMAP dimensionality reduction with  $n\_neighbors=250$ . (b) UMAP embedding of articles with color indicating the source (AF, purple; arXiv, green). (c) UMAP embedding of articles with color indicating date of publication. Arrows superimposed to indicate direction of temporal evolution. (d) UMAP embedding of articles with color indicating cluster membership as determined with k-means ( $k=5$ ). Inset shows sum of residuals as a function of clusters  $k$ , with an arrow highlighting the chosen number of clusters.

We were curious to see if the five clusters identified by k-means map onto existing distinctions in the field. When identifying the most prolific authors in each cluster, we noticed strong differences<sup>[3]</sup> ([consistent with previous work](#)) that suggests that author identity is an important indicator of research direction).

cluster 1; $N = 567$ (agent alignment)	cluster 2; $N = 988$ (alignment foundations)	cluster 3; $N = 593$ (tool alignment)	cluster 4; $N = 383$ (AI governance)	cluster 5; $N = 670$ (value alignment)
S. Levine (55)	S. Armstrong (154)	J. Steinhardt (20)	D. Kokotajlo (21)	S. Armstrong (54)
P. Abbeel (34)	S. Garrabrant (95)	D. Hendrycks (17)	A. Dafoe (19)	S. Byrnes (32)
A. Dragan (29)	A. Demski (94)	E. Hubinger (14)	G. Worley III (11)	P. Christiano (29)
S. Russell (23)	J. Wentworth (57)	P. Christiano (13)	J. Clarck (10)	R. Ngo (25)
S. Armstrong (22)	"Diffractor" (44)	P. Kohli (11)	S. Armstrong (9)	R. Shah (25)

**Table 2: Researchers with the highest number of articles per cluster.**

Clusters as determined in Fig. 2, with number of articles per cluster  $N$ . Number in brackets behind researcher name indicates number of articles published by that researcher. Note: "Diffractor" is an undisclosed pseudonym.

By skimming articles in each cluster and given the typical research published by the authors, we suggest the following putative descriptions of each cluster:

1. **cluster one:** *Agent alignment* is concerned with the problem of aligning agentic systems, i.e. those where an AI performs actions in an environment and is typically trained via reinforcement learning.
2. **cluster two:** *Alignment foundations* research is concerned with *deconfusion* research, i.e. the task of establishing formal and robust conceptual foundations for current and future AI Alignment research.
3. **cluster three:** *Tool alignment* is concerned with the problem of aligning non-agentic (tool) systems, i.e. those where an AI transforms a given input into an output. The

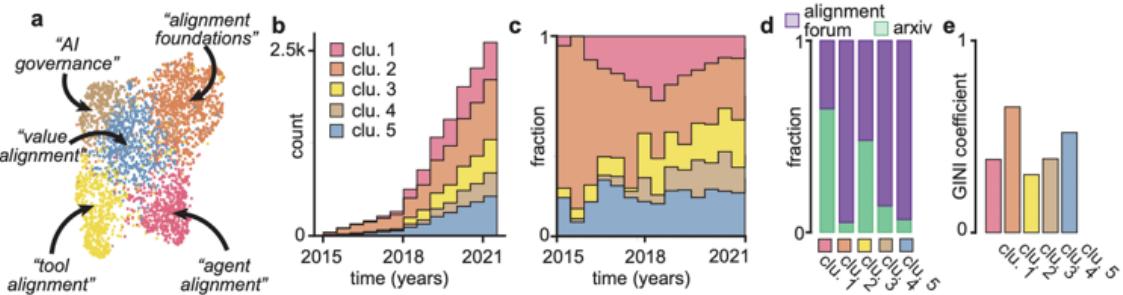
current, prototypical example of tool AIs is the "large language model".

4. **cluster four:** *AI governance* is concerned with how humanity can best navigate the transition to advanced AI systems. This includes focusing on the political, economic, military, governance, and ethical dimensions.
5. **cluster five:** *Value alignment* is concerned with understanding and extracting human preferences and designing methods that stop AI systems from acting against these preferences.

We note that **these descriptions are chosen to be descriptive, not prescriptive**. Our approach has the advantage of being (comparatively<sup>[4]</sup>) unbiased and can therefore serve as a baseline against which other (more prescriptive) descriptions of the landscape can be compared ([Krakovna's paradigms](#), [FLI landscape](#), [Christiano's landscape](#), [Nanda's overview](#), ...). Discrepancies between these descriptions and ours can serve as important information for funding agencies (to identify neglected areas) and AI Governance researchers (for early identification of natural categories for regulation).

## Research dynamics vary across the identified clusters

We further note some properties of the identified clusters (Fig. 3a). The cluster labeled as "alignment foundations" contains most of the seminal work in the field (Fig. 3b,c), but remains largely disconnected from the more applied "agent alignment" and "tool alignment" research (Fig. 3a). Furthermore, most "alignment foundations" work is published on the Alignment Forum (Fig. 3d) and it has the largest inequality in terms of "number of articles per researcher" (Fig. 3e). This corroborates an observation that was made before: **While critically important, alignment foundations research appears to be poorly integrated into more applied alignment research, and the research remains insular and pushed by comparatively few researchers.**

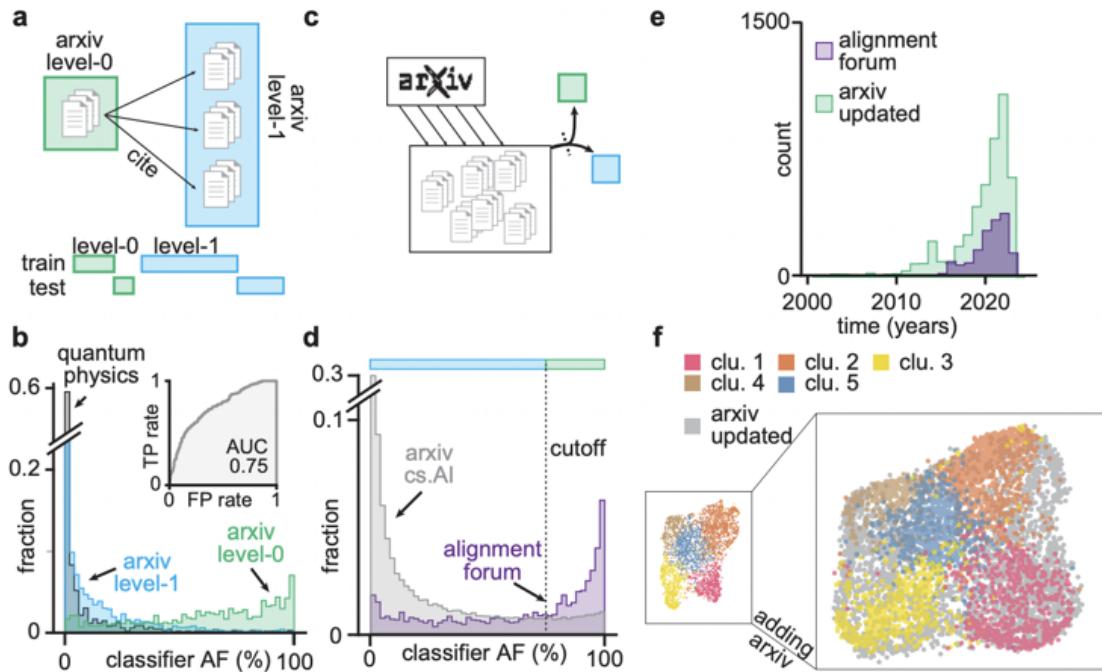


**Figure 3: Characteristics of research clusters corroborate potential usefulness of decomposition.** (a) UMAP embedding of articles with color indicating cluster membership as in Fig. 2d. Labels assigned to each cluster are putative descriptions of a common research focus across articles in the cluster. (b) Number of articles published per year, colored by cluster membership. (c) Fraction of articles published by cluster membership as a function of time. (d) Fraction of articles from AF or arXiv as a function of cluster membership. (e) GINI inequality coefficient of articles per researcher as a function of article cluster membership.

## Leveraging dataset to train an AI alignment research classifier

After having identified the five clusters, we returned to the issue we noted at the onset of our analysis: the apparent decrease in publications on the arXiv in recent years (Fig. 1a). We were skeptical about this and hypothesized that our data collection might have missed relevant recent articles<sup>[5]</sup>. Therefore, we trained a logistic regression classifier to distinguish alignment articles (level-0) from articles cited by alignment articles (level-1) (Fig. 4 a). The resulting classifier achieved good performance and generalized well to papers from unrelated sources (Fig. 4 b). We then scraped all the articles from the arXiv cs.AI category and asked our classifier to score them (Fig. 4c,d). Based on the distribution of scores of Alignment Forum posts (Fig. 4d) and after skimming the relevant articles, we chose a threshold of 75% as a reasonable trade-off between false positives and false negatives.

When adding the arXiv articles above the cutoff to our dataset, we observed a rapid increase in publications also on the arXiv (Fig. 4e). To test if our clustering is robust to this increase, we repeated the UMAP projection with the updated dataset and found that, indeed, the clusters are still in distinct regions of the manifold (Fig. 4f). Interestingly, the added literature appears to fill some of the gaps between "alignment foundations" and "agent alignment" research.



**Figure 4: An AI alignment research classifier for filtering new publications.** (a) Illustration of arXiv level-0 articles (alignment research; green) and level-1 articles (cited alignment research articles; blue). Bottom: Schematic of test-train split (20%-80% for training of a logistic regression classifier). (b) Fraction of articles as a function of classifier score for arXiv level-0 (green), level-1 (blue), and arXiv articles on quantum physics (grey). (c) Illustration of procedure for filtering arXiv articles. After querying articles from the cs.. section of arXiv, the logistic regression classifier assigns a score between 0 and 1. (d) Fraction of articles as a function of classifier score for articles from the cs.AI section of arXiv (grey) and AF (purple). Dashed line indicates cutoff for classifying articles as arXiv level-0 (75%). (e) Number of articles published as a function of time on AF (purple) and arXiv (green), according to the cutoff in panel d. (f) Left inset: Original UMAP embedding from Fig. 2d. Right: UMAP embedding of all original articles and updated arXiv articles with color indicating cluster membership as in Fig. 2d or that the article is filtered from the arXiv (g)

## Closing remarks

The primary output from our project is the curated dataset of alignment research articles. We hope the dataset might serve as the basis for

- a semantic search service that returns relevant literature (see prototype [here](#)).
- [writing assistants in the form of fine-tuned large-language models](#).
- [projects to preserve AI Safety research in case of catastrophic events](#).

If you have other ideas for how to use the dataset, please don't hesitate to reach out to us; we're excited to help.

Furthermore, we hope that the secondary outcome from our project (the analysis in this post) can aid both funding agencies and new researchers entering the field to orient themselves and contextualize the research.

As we plan to continue this line of research, we are happy about any and all feedback on the dataset and the analysis, as well as hints and pointers about things we might have missed.

*Acknowledgments: We thank Daniel Clothiaux for help with writing the code and extracting articles. We thank Remmelt Ellen, Adam Shimi, and Arush Tagade for feedback on the research. We thank Chu Chen, Ömer Faruk Sen, Hey, Nihal Mohan Moodbidri, and Trinity Smith for cleaning the audio transcripts.*

1. ^

We will make some finishing touches on the repository over the next few weeks after this post is published.

2. ^

Kirchner, J. H., Smith, L., Thibodeau, J., McDonnell, K., and Reynolds, L. "Understanding AI alignment research: A Systematic Analysis." *arXiv preprint arXiv:2206.02841* (2022).

3. ^

Except for Stuart Armstrong, who publishes prolifically across all clusters.

4. ^

Remaining biases include:

- differences in formatting between arxiv and AF articles that bias the embedding
- some (important) topics might not have any documentation due to infohazards
- by implicitly focusing on number of published articles (rather than f.e. the "volume occupied in semantic space") we bias our analysis in favor of questions that can be written about more easily

5. ^

We took the [TAI Safety Bibliographic Database](#) from early 2020 as a starting point and manually added relevant articles from other existing bibliographies or based on our judgment. We were very conservative in this step, as we wanted to make sure that our dataset includes as few false positives as possible.

# Limits to Legibility

From time to time, someone makes the case for why transparency in reasoning is important. The latest conceptualization is [Epistemic Legibility](#) by Elizabeth, but the core concept is similar to [reasoning transparency](#) used by OpenPhil, and also has some similarity to [A Sketch of Good Communication](#) by Ben Pace.

I'd like to offer a gentle pushback. The tl;dr is in [my comment](#) on Ben's post, but it seems useful enough for a standalone post.

*"How odd I can have all this inside me and to you it's just words."* — David Foster Wallace

## When and why reasoning legibility is hard

Say you demand transparent reasoning from AlphaGo. The [algorithm](#) has roughly two parts: tree search and a neural network. Tree search reasoning is naturally legible: the "argument" is simply a sequence of board states. In contrast, the neural network is mostly illegible - its output is a figurative "feeling" about how promising a position is, but that feeling depends on the aggregate experience of a huge number of games, and it is extremely difficult to explain transparently how a particular feeling depends on particular past experiences. So AlphaGo would be able to present part of its reasoning to you, but not the most important part.<sup>[1]</sup>

Human reasoning uses both: cognition similar to tree search (where the steps can be described, written down, and explained to someone else) *and* processes not amenable to introspection (which function essentially as a black box that produces a "feeling"). People sometimes call these latter signals "intuition", "implicit knowledge", "taste", "S1 reasoning" and the like. Explicit reasoning often rides on top of this.

Extending the machine learning metaphor, the problem with *human interpretability* is that "mastery" in a field often consists precisely in having some well-trained black box neural network that performs fairly opaque background computations.

## Bad things can happen when you demand explanations from black boxes

The second thesis is that it often makes sense to assume the mind runs distinct computational processes: one that actually makes decisions and reaches conclusions, and [another](#) that produces justifications and rationalizations.

In my experience, if you have good introspective access to your own reasoning, you may occasionally notice that a conclusion C depends mainly on some black box, but at the same time, you generated a plausible legible argument A for the same conclusion **after you reached the conclusion C**.

If you try running, say, [Double Crux](#) over such situations, you'll notice that even if someone refutes the explicit reasoning A, you won't quite change the conclusion to  $\neg C$ . The legible argument A was not the real crux. It is quite often the case that (A) is essentially fake (or low-weight), whereas the black box is hiding a reality-tracking model.

Stretching the AlphaGo metaphor a bit: AlphaGo could be easily modified to find a few specific game "rollouts" that turned out to "explain" the mysterious signal from the neural network. Using tree search, it would produce a few specific examples how such a position may evolve, which would be selected to agree with the neural net prediction. If AlphaGo showed them to you, it might convince you! But you would get a completely superficial understanding of why it evaluates the situation the way it does, or why it makes certain moves.

## Risks from the legibility norm

When you make a strong norm pushing for too straightforward "epistemic legibility", you risk several bad things:

First, you increase the pressure on the "justification generator" to mask various black boxes by generating arguments supporting their conclusions.

Second, you make individual people dumber. Imagine asking a Go grandmaster to transparently justify his moves to you, and to play the moves that are best justified - if he tries to play that way, he will become a much weaker player. A similar thing applies to AlphaGo - if you allocate computational resources in such a way that a much larger fraction is consumed by tree search at each position, and less of the neural network is used overall, you will get worse outputs.

Third, there's a risk that people get convinced based on bad arguments - because their "justification generator" generated a weak legible explanation, you managed to refute it, and they updated. The problem comes if this involves discarding the output of the neural network, which was much smarter than the reasoning they accepted.

## What we can do about it

My personal impression is that society as a whole would benefit from more transparent reasoning on the margin.

What I'm not convinced of, at all, is that trying to reason much more transparently is a good goal for aspiring rationalists, or that some naive (but memetically fit) norms around epistemic legibility should spread.

To me, it makes sense for some people to specialize in very transparent reasoning. On the other hand, it also makes sense for some people to mostly "try to be better at Go", because legibility has various hidden costs.

A version of transparency that seems more robustly good to me is the one that takes legibility to a meta level. It's perfectly fine to refer to various non-interpretable processes and structures, but we should ideally add a description of what data they

are trained on (e.g. "I played at the national level"). At the same time, if such black-box models outperform legible reasoning, it should be considered fine and virtuous to use models which work. You should play to win, if you can.

## Examples

An example of a common non-legible communication:

**A:** Can you explain why you feel that getting this person to implement a "Getting Things Done" system is not a good idea?

**B:** I don't know exactly, I feel it won't do him any good

An example of how to make the same conversation worse by naive optimization for legibility

**A:** Can you explain why you feel that getting this person to implement a "Getting Things Done" system is not a good idea?

**B:** I read a thread on Twitter yesterday where someone explained that research on similar motivational techniques does not replicate, and also another thread where someone referenced research that people who over-organize their lives are less creative.

**A:** Those studies are pretty weak though.

**B:** Ah I guess you're right.

An example of how to actually improve the same conversation by striving for legibility:

**A:** Can you explain why you feel that getting this person to implement a "Getting Things Done" system is not a good idea?

**B:** I guess I can't explain it transparently to you. My model of this person just tells me that there is a fairly high risk that teaching them GTD won't have good results. I think it's based on experience with a hundred people I've met on various courses who are trying to have a positive impact on the world. Also, when I had similar feelings in the past, it turned out they were predictive in more than half of the cases.

If you've always understood the terms "reasoning transparency" or "epistemic legitimacy" in the spirit of the third conversation, and your epistemology routinely involves steps like "*I'm going to trust this black-box trained on lots of data a lot more than this transparent analysis based on published research*", then you're probably safe.

## How this looks in practice

In my view, it is pretty clear that some of the main cruxes of current disagreements about AI alignment are beyond the limits of legible reasoning. (The current limits, anyway.)

In my view, some of these intuitions have roughly the "black-box" form explained above. If you try to understand the disagreements between e.g. Paul Christiano and Eliezer Yudkowsky, you often end up in a situation where the real difference is "taste", which influences how much weight they give to arguments, how good or bad various future "board positions" are evaluated to be, etc. Both Elizer and Paul are extremely smart, have spent more than a decade thinking about AI safety and even more time on relevant topics such as ML or decision theory or epistemics.

A person new to AI safety evaluating their arguments is roughly at a similar position to a Go novice trying to make sense of two Go grandmasters disagreeing about a board, with the further unfortunate feature that you can't just make them play against each other, because in some sense they are both playing for the same side.

This isn't a great position to be in. But in my view it's better to understand where you are rather than, for example, naively updating on a few cherry-picked rollouts.

## See also

- [The decision version](#)

*Thanks to Gavin for help with writing this post.*

1. [^](#)

We can go even further if we note that the later AlphaZero policy network doesn't use tree search when playing.

# Let's See You Write That Corrigibility Tag

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

The [top-rated comment](#) on "[AGI Ruin: A List of Lethalities](#)" claims that many other people could've written a list like that.

"Why didn't you challenge anybody else to write up a list like that, if you wanted to make a point of nobody else being able to write it?" I was asked.

Because I don't actually think it does any good, or persuades anyone of anything, people don't like tests like that, and I don't really believe in them myself either. I couldn't pass a test somebody else invented around something *they* found easy to do, for many such possible tests.

But people asked, so, fine, let's actually try it this time. Maybe I'm wrong about how bad things are, and will be pleasantly surprised. If I'm never pleasantly surprised then I'm obviously not being pessimistic enough yet.

So: As part of my current fiction-writing project, I'm currently writing a list of some principles that dath ilan's Basement-of-the-World project has invented for describing AGI corrigibility - the sort of principles you'd build into a Bounded Thing meant to carry out some single task or task-class and not destroy the world by doing it.

So far as I know, every principle of this kind, except for Jessica Taylor's "quantilization", and "myopia" (not sure who correctly named this as a corrigibility principle), was invented by myself; eg "low impact", "shutdownability". (Though I don't particularly think it hopeful if you claim that somebody else has publication priority on "low impact" or whatevs, in some stretched or even nonstretched way; ideas on the level of "low impact" have always seemed cheap to me to propose, harder to solve before the world ends.)

Some of the items on dath ilan's upcoming list out of my personal glowfic writing have already been written up more seriously by me. Some haven't.

I'm writing this in one afternoon as one tag in my cowritten online novel about a dath ilani who landed in a D&D country run by Hell. ~~One and a half thousand words or so, maybe.~~ (2169 words.)

How about you try to do better than the tag overall, before I publish it, upon the topic of corrigibility principles on the level of "myopia" for AGI? It'll get published in a day or so, possibly later, but *I'm* not going to be spending more than an hour or two polishing it.

# Leaving Google, Joining the Nucleic Acid Observatory

In 2017 I [rejoined Google](#) to earn money to donate. At the time I thought earning to give was probably not where I could have the most impact, but I [wasn't able to find](#) other options that were a good fit for me personally. Over the last five years a few things have changed:

- There is [substantially more funding available within effective altruism](#), and so the importance of [earning to give](#) has continued to decrease relative to doing things that aren't mediated by donations.
- The place where I've [long been](#) most skeptical of the value of work to reduce existential risk is the [lack of good feedback loops](#). There are now several major areas, however, where it seems pretty practical to tell whether you're making good progress and executing well. I'm [especially enthusiastic](#) about [concrete projects](#) in avoiding or mitigating [catastrophic pandemics and other biological risks](#).
- I've found an in-person role in Boston where I can apply my skills to one of these relatively tractable areas of existential risk reduction.

So: today will be my last day at Google, and Monday will be my first day at the [Nucleic Acid Observatory](#) (NAO). We'll be building a system to collect wastewater samples and sequence their nucleic acids, with the goal of catching potential future pandemics earlier. More details in the [EA Forum post](#) and much more in [the paper](#).

In looking for things that I might do instead of earning to give, I identified several other strong candidates for ways to apply software engineering skills to making the world better. If you're thinking of making a similar move, let me know and I'd be happy to give you an overview of what I found and potentially give introductions. While normally I prefer people [comment publicly instead of sending private messages](#), this is the kind of thing where I'm happy to receive messages.

While I'm overall [quite mixed](#) on how the increased focus on applying your career has made EA more demanding, in my particular case I think it's pushed me in a good direction.

Timeline of this decision:

- Weekend of 2022-03-26: Informal in-person discussion with EA friends I haven't seen in a while gets me thinking again about moving into something more directly useful.
- 2022-03-28: One of these friends, who also happens to work at [80,000 Hours](#), follows up by email and gets me thinking specifically about how bio could offer a good combination of impactful, in-person, and in-Boston.
- 2022-03-31: I [write](#) "since now that there is so much more money available in the EA movement I'm back to thinking about doing something other than earning to give".

- 2022-04-04: I write to [Will Bradshaw](#) to see if he has ideas about where I might be helpful, though for travel and personal reasons we don't end up meeting to talk until 2022-04-29.
- 2022-04-08: I write to [Chris Bakerlee](#) at the Open Philanthropy project, who gives good advice and suggestions of people to talk to.
- 2022-04-13 through 2022-04-27: In the UK, watching the kids while Julia attends [EA Global](#) and then visiting EA friends there after. Talked to quite a few different people about options here.
- 2022-05: A lot of reading from the two lists ([Greg](#)'s and [Chris'](#)) that [80,000 Hours](#) links.
- 2022-05: Talking to three different groups I was strongly considering joining. In addition to the NAO this was [Alvea](#) and [SecureDNA](#), both of which I think highly of.
- 2022-05-23: At dinner with housemates I realize that with three strong options and several other ideas for things that I might do if those fell through that I'm very unlikely to stay at Google.
- 2022-05-24: Gave notice to my manager. I'm out on leave this week, though, so I don't start handoffs yet.
- 2022-05-31: Announce to my team that I'm leaving and start handing things off.
- 2022-06-04: Decided to join the NAO.
- 2022-06-10: Last day at Google.
- 2022-06-13: First day at the NAO.

As [last time](#), I'm pretty sad to be leaving Google. It's been a great place to work, and I especially like my team and the [work we do](#). I've [built up](#) a deep understanding of the web platform and advertising ecosystem, and while this domain knowledge isn't especially altruistically useful it's been fascinating and challenging work. I'll still be following the progress of [subresource bundles](#) and the [privacy sandbox APIs](#), and I'm going to miss so many people!

In writing my goodbye emails I saw that this time I've been at Google for 1,739 days, thirteen shy of my previous record, 1,752. While I don't think I would have stayed another two weeks just to make it even, I'm glad this didn't occur to me until after I'd given notice.

*Comment via: [facebook](#)*

# "Pivotal Acts" means something specific

The term [Pivotal Act](#) was written up on Arbilal in 2015. I only started hearing it discussed in 2020, and then it rapidly started seeing more traction when the [MIRI 2021 Conversations](#) were released.

I think people mostly learned the word via context-clues, and never actually read the article.

I have some complaints about how Eliezer defined the term, but I think if you're arguing with MIRI-cluster people and haven't [read the article in full](#) you may have a confusing time.

The arbilal page for Pivotal Act begins:

The term 'pivotal act' in the context of [AI alignment theory](#) is a [guarded term](#) to refer to actions that will make a large positive difference **a billion years later**.

And then, almost immediately afterwards, the article goes on to reiterate "this is a guarded term", and explains why. i.e. this is a jargon term that people are going to be *very tempted* to stretch the definition of, but which it's really important not to stretch the definition of.

The article notes:

## Reason for guardedness

[Guarded definitions](#) are deployed where there is reason to suspect that a concept will otherwise be over-extended. The case for having a guarded definition of 'pivotal act' (and another for 'existential catastrophe') is that, after it's been shown that event X is maybe not as important as originally thought, one side of that debate may be strongly tempted to go on arguing that, wait, really it could be "relevant" (by some [strained](#) line of possibility).

It includes a bunch of examples (you really should go read the [full article](#)), and then notes:

**Discussion:** Many [strained arguments](#) for X being a pivotal act have a step where X is an input into a large pool of goodness that also has many other inputs. A ZF provability oracle would advance mathematics, and mathematics can be useful for alignment research, but there's nothing obviously game-changing about a ZF oracle that's specialized for advancing alignment work, and it's unlikely that the effect on win probabilities would be large relative to the many other inputs into total mathematical progress.

Similarly, handling trucker disemployment would only be one factor among many in world economic growth.

By contrast, a genie that uploaded human researchers putatively would *not* be producing merely one upload among many; it would be producing the only

uploads where the default was otherwise no uploads. In turn, these uploads could do decades or centuries of unrushed serial research on the AI alignment problem, where the alternative was rushed research over much shorter timespans; and this can plausibly make the difference by itself between an AI that achieves ~100% of value versus an AI that achieves ~0% of value. At the end of the extrapolation where we ask what difference everything is supposed to make, we find a series of direct impacts producing events qualitatively different from the default, ending in a huge percentage difference in how much of all possible value gets achieved.

By having narrow and guarded definitions of 'pivotal acts' and 'existential catastrophes', we can avoid bait-and-switch arguments for the importance of research proposals, where the 'bait' is raising the apparent importance of 'AI safety' by discussing things with large direct impacts on astronomical stakes (like a paperclip maximizer or Friendly sovereign) and the 'switch' is to working on problems of dubious astronomical impact that are inputs into large pools with many other inputs.

I see people stretching "pivotal act" to mean "things that delay AGI for a few years or decades", which isn't what the term is meant to mean.

[Full article here.](#)

# Will Capabilities Generalise More?

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

[Nate](#) and [Eliezer](#) (Lethality 21) claim that capabilities generalise further than alignment once capabilities start generalising far at all. However, they have not articulated particularly detailed arguments for why this is the case. In this post I collect the arguments for and against the position I have been able to find or generate, and develop them (with a few hours' effort). I invite you to join me in better understanding this claim and its veracity by contributing your own arguments and improving mine.

*Thanks to these people for their help with writing and/or contributing arguments: Vikrant Varma, Vika Krakovna, Mary Phuong, Rory Grieg, Tim Genewein, Rohin Shah.*

## For:

### **1. Capabilities have much shorter description length than alignment.**

There are simple “laws of intelligence” that underwrite highly general and competent cognitive abilities, but no such simple laws of corrigibility or laws of “doing what the principal means” – or at least, any specification of these latter things will have a higher description length than the laws of intelligence. As a result, most R&D pathways optimising for capabilities and alignment with anything like a simplicity prior ([for example](#)) will encounter good approximations of general intelligence earlier than good approximations of corrigibility or alignment.

### **2. Feedback on capabilities is more consistent and reliable than on alignment.**

Reality hits back on cognitive strategies implementing capabilities – such as forming and maintaining accurate beliefs, or making good predictions – more consistently and reliably than any training process hits back on motivational systems orienting around incorrect optimisation targets. Therefore there’s stronger outer optimisation pressure towards good (robust) capabilities than alignment, so we see strong and general capabilities first.

### **3. There’s essentially only one way to get general capabilities and it has a free parameter for the optimisation target.**

There are many paths but only one destination when it comes to designing (via optimisation) a system with strong capabilities. But what those capabilities end up

being directed at is path- and prior-dependent in a way we currently do not understand nor have much control over.

## **4. Corrigibility is conceptually in tension with capability, so corrigibility will fail to generalise when capability generalises well.**

Plans that actually work in difficult domains need to preempt or adapt to obstacles. Attempts to steer or correct the target of actually-working planning are a form of obstacle, so we would expect capable planning to resist correction, limiting the extent to which alignment can generalise when capability starts to generalise.

## **5. Empirical evidence: human intelligence generalised far without staying aligned with its optimisation target.**

There is empirical/historical support for capabilities generalising further than alignment to the extent that the analogy of AI development to the evolution of intelligence holds up.

## **6. Empirical evidence: goal misgeneralisation happens.**

There is weak empirical support for capabilities generalising further than alignment in the fact that it is possible to create demos of goal misgeneralisation (e.g., <https://arxiv.org/abs/2105.14111>).

## **7. The world is simple whereas the target is not.**

There are relatively simple laws governing how the world works, for the purposes of predicting and controlling it, compared to the principles underlying what humans value or the processes by which we figure out what is good. (This is similar to For#1 but focused on knowledge instead of cognitive abilities.) (This is in direct opposition to Against#3.)

## **8. Much more effort will be poured into capabilities (and $d(\text{progress})/d(\text{effort})$ for alignment is not so much higher than for capabilities to counteract this).**

We'll assume alignment is harder based on the other arguments. For why more effort will be put into capabilities, there are two economic arguments: (a) at lower capability levels there is more profitability in advancing capabilities than alignment specifically, and (b) data about reality in general is cheaper and more abundant than data about any particular alignment target (e.g., human-preference data).

This argument is similar to For#2 but focused more on the incentives faced by R&D organisations and efforts: paths to developing capabilities are more salient and attractive.

## **9. Alignment techniques will be shallow and won't withstand the transition to strong capabilities.**

There are two reasons: (a) we don't have a principled understanding of alignment and (b) we won't have a chance to refine our techniques in the strong capabilities regime.

If advances in a core of general reasoning cause performance on specific domains like bioengineering or psychology to look "jumpy", this will likely happen at the same time as a jump in the ability to understand and deceive the training process, and evade the shallow alignment techniques.

### **Against:**

#### **1. Optimal capabilities are computationally intractable; tractable capabilities are more alignable.**

For example, it may be that the structure of the cognition of tractable capabilities does not look like optimal planning - there's no obvious factorisation into goals and capabilities. Convergent instrumental subgoals may not apply strongly to the intelligences we actually find.

#### **2. Reality hits back on the models we train via loss functions based on reality-generated data. But alignment also hits back on models we train, because we also use loss functions (based on preference data). These seem to be symmetrically powerful forces.**

In fact we care a lot about models that are deceptive or harmful in non-x-risky ways, and spend massive effort curating datasets that describe safe behaviour. As models get more powerful, we will be able to automate the process of generating better

datasets, including through AI assistance. Eventually we will effectively be able to constrain the behaviour of superhuman systems with the sheer quantity and diversity of training data.

### **3. Alignment only requires building a pointer, whereas capability requires lots of knowledge. Thus the overhead of alignment is small, and can ride increasing capabilities.**

(Example of a similar structure, which gives some empirical evidence: millions of dollars to train GPT-3 but only thousands of dollars to finetune on summarisation.)

### **4. We may have schemes for directing capabilities at the problem of oversight, thus piggy-backing on capability generalisation.**

E.g. debate and recursive reward modelling. Furthermore, overseers are asymmetrically advantaged (e.g. because of white-box access or the ability to test in simulation on hypotheticals).

### **5. Empirical evidence: some capabilities improvements have included corresponding improvements in alignment.**

It has proved possible, for example fine-tuning language models on human instructions, to build on capabilities to advance alignment. Extrapolating from this, we might expect alignment to generalise alongside capabilities. For example, billions of tokens are required for decent language capabilities but then only thousands of human feedback points are required to point them at a task.

### **6. Capabilities might be possible without goal-directedness.**

Humans are arguably not strongly goal-directed. We seem to care about lots of different things, and mostly don't end up with a desire to strongly optimise the world towards a simple objective.

Also, we can build tool AIs (such as a physics simulator or a chip designer) which are targeted at such a narrow domain that goal-directedness is not relevant since they aren't strategically located in our world. These AIs are valuable enough to produce economic bounties while coordinating against goal-directed AI development.

## **7. You don't actually get sharp capability jumps in relevant domains**

The AI industry will optimise hard on all economically relevant domains (like bioengineering, psychology, or AI research), which will eliminate capability overhangs and cause progress on these domains to look smooth. This means we get to test our alignment techniques on slightly weaker AIs before we have to rely on them for slightly stronger AIs. This will give us time to refine them into deep alignment techniques rather than shallow ones, which generalise enough.

# CFAR Handbook: Introduction

The [Center for Applied Rationality](#) is a Bay Area non-profit that, among other things, ran lots of workshops to offer people tools and techniques for solving problems and improving their thinking. Those workshops were accompanied by a reference handbook, which has been available as a [PDF](#) since 2020.

The handbook hasn't been substantially updated since it was written in 2016, but it remains a fairly straightforward primer for a lot of core rationality content. The LW team, working with the handbook's author [Duncan Sabien](#), have decided to republish it as a lightly-edited sequence, so that each section can be linked on its own.

In the workshop context, the handbook was a *supplement* to lectures, activities, and conversations taking place between participants and staff. Care was taken to emphasize the fact that each tool or technique or perspective was only as good as it was effectively applied to one's problems, plans, and goals. The workshop was intentionally structured to cause participants to actually try things (including iterating on or developing their own versions of what they were being shown), rather than simply passively absorb content. Keep this in mind as you read—mere knowledge of how to exercise does not confer the benefits of exercise!

Discussion is strongly encouraged, and disagreement and debate are explicitly welcomed. Many LWers (including the staff of CFAR itself) have been tinkering with these concepts for years, and will have developed new perspectives on them, or interesting objections to them, or thoughts about how they work or break in practice. What follows is a historical artifact—the rough state-of-the-art at the time the handbook was written, circa 2017. That's an excellent jumping-off point, especially for newcomers, but there's been a lot of scattered progress since then, and we hope some of it will make its way into the comments.

# Who models the models that model models? An exploration of GPT-3's in-context model fitting ability

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

## Introduction

Much has been written and much has been observed about the abilities of GPT-3 on [many tasks](#). Most of these capabilities, though not all, pertain to **writing convincing text**, but, not to undermine GPT-3's impressiveness at performing these tasks, we might call this the *predictable* part of its oeuvre. It only makes sense that a better language modelling is, well, going to be better at writing text.

Deeper and comparatively much less explored is the *unpredictable* ability of GPT-3 to learn new tasks by just **seeing a few examples**, without any training/backpropagation – the so called in-context learning (sometimes called metalearning).

The [original paper announcing GPT-3](#) contained a handful of examples (perhaps mostly notably examples of GPT-3 learning to perform arithmetic, e.g. accurate addition of up to 5-digit numbers), Gwern has also insightfully [written about it](#), Frieda Rong has performed [some interesting experiments](#), and there have been various [other experiments](#) one could chance upon. My curiosity being piqued but not sated by these experiments, and also having had the feeling that, as captivating the arithmetic examples were, they weren't the most *natural* question one could ask about a stochastic model's quantitative capabilities – I decided to investigate whether GPT-3 could *fit numerical models in-context*.

## The Setup

What does it mean to fit a model in-context? Well, recall that GPT-3 has a context window (roughly speaking: text it considers before generating additional text) of length 2048 tokens (roughly speaking: (parts of the) words, punctuation, etc.) so the idea is to put feature vectors *inside* that context window. Of course, this means you cannot fit any larger or higher-dimensional dataset in there.<sup>[1]</sup>

In practice this means prompting GPT-3 on input like:

```
Input: 94, 47, 84, 31, output = 2
Input: 89, 51, 73, 31, output = 1
[...]
Input: 96, 51, 80, 38, output = 2
Input: 90, 37, 76, 27, output =
```

And then taking its output, i.e. the text it generates, as the prediction.

A couple more technical details: In all the experiments I performed, all the numbers were integers. The GPT-3's outputs were always sampled with temperature 0, i.e. they were deterministic – only the most probable output was considered. I restricted myself that way for simplicity, but hopefully some future work looks at the full distribution over outputs.

(In the rest of this post I share GPT-3's results without too much warning, so if you'd like to make your own predictions about how GPT-3 does at simple classification and regression tasks in-context, now would be the time to **pause reading and make your general predictions.**)

## Iris Dataset

Just as there is the [MNIST dataset](#) for visual recognition, so too there is a small low-dimensional classification dataset which every would-be classifier has to solve or else sink – the [Iris dataset](#), composed of 150 observations of sepal/petal height/width of three species of iris (50 observations each). Note that classification of the Iris dataset is in some sense trivial, with simple algorithms achieving near-perfect accuracy, but a model still needs to do some meaningful learning, given that it

needs to learn to differentiate between three classes based on just a few dozen four-dimensional vectors.

In order to prevent leakage, as the Iris dataset is all over the internet, as well as to get the more-easily-palatable integers out, I transformed every feature with the transformation

$$x_{\text{new}} = \text{round}(14x_{\text{old}} + 6).$$

I also split the dataset into 50% train and 50% test folds – so that the model "trained" on, or rather *looked* at 75 examples.

I hadn't quite been expecting what had followed – I had expected GPT-3 to catch on to some patterns, to have a serviceable but not-quite-impressive accuracy – instead, the accuracies, averaged over 5 random dataset shufflings, for Ada (350M params), Babbage (1.3B), Curie (6.7B), and Davinci (175B), compared to kNN (k=5) and logistic regression were:

Model	Accuracy
kNN	95.73%
Logistic regr.	96.26%
Ada	89.86%
Babbage	93.06%
Curie	95.20%
Davinci	95.73%

So Curie and Davinci did about as well as kNN[\[2\]](#) and logistic regression. GPT-3, just by *looking* at feature vectors, solves Iris.

I also conducted a few other experiments with the Iris dataset. One was not labelling the "input" or "output", but just sending bare numbers, like this:

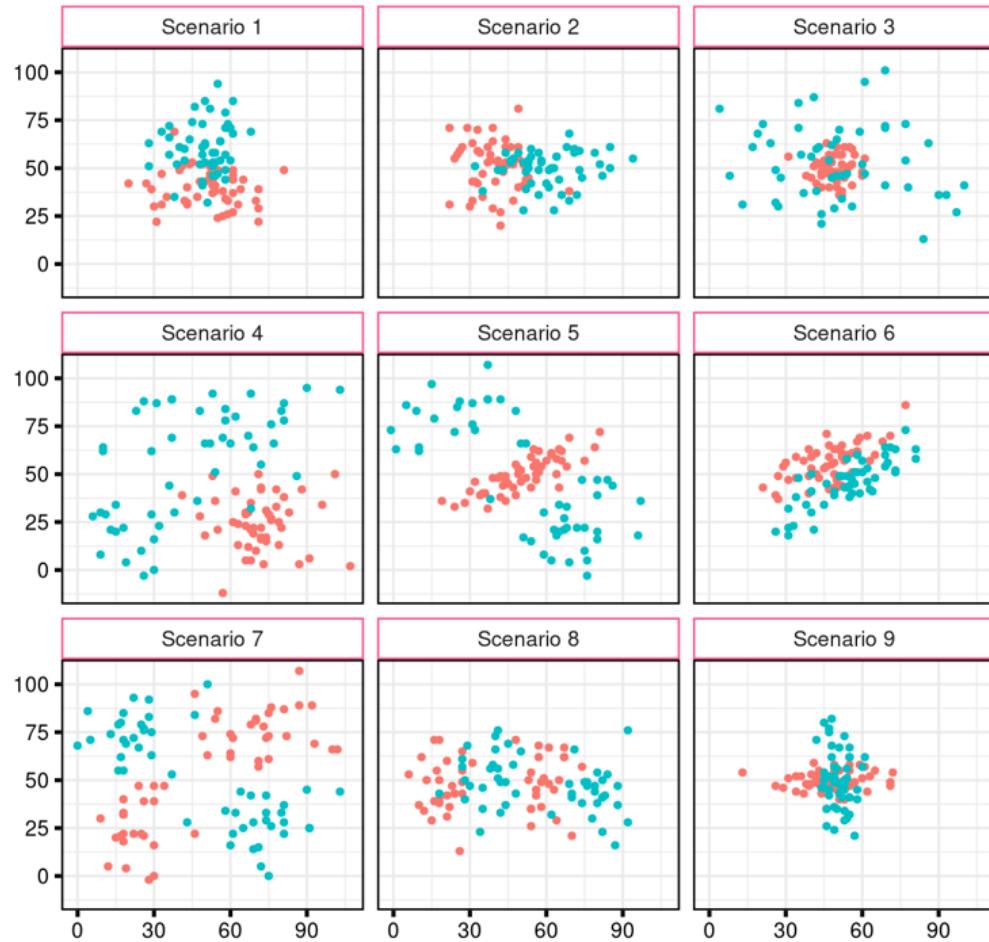
```
94, 47, 84, 31, 2
89, 51, 73, 31, 1
[...]
91, 48, 75, 31, 2
96, 51, 80, 38,
```

This seemed to degrade the results very slightly. Scaling up numbers so that all features were in the hundreds also seemed to potentially degrade performance by a few percentage points, but I didn't investigate that exhaustively either.[\[3\]](#)

## 2D binary classification, generally

These results, while interesting, only show GPT-3's model-fitting ability on one dataset. It might, indeed, be that the Iris dataset is, for some reason, unusually easy to classify for GPT-3. That's why I conducted some more experiments, in a lower dimension, where I could try out a lot different things.

To the end of trying out a lot of things, I constructed 9 "typical binary classification scenarios" – I tried to think up a set of class distributions which would capture a decent number of realistic-looking two-class binary classification cases, as well as some slightly adversarial ones – e.g. scenario 7 is the analogue of the "XOR" problem, which perceptrons famously cannot learn because they define a linear boundary. Below you can see a random sample of each of these scenarios.



For each scenario I sampled a dataset three times; each of those datasets had 50 "train" examples and 30 test examples. Here are the results, comparing between kNN ( $k=5$ <sup>[4]</sup>), logistic regression, a custom text-based classifier which I wrote thinking about the easiest text-based-algorithm GPT could learn<sup>[5]</sup>, and GPT-3.

Model	Scen. 1	Scen. 2	Scen. 3	Scen. 4	Scen. 5	Scen. 6	Scen. 7	Scen. 8	Scen. 9
kNN	75.56%	78.89%	71.11%	93.33%	98.89%	75.56%	90.0%	83.33%	68.89%
Logistic regr.	75.56%	78.89%	46.67%	93.33%	38.89%	81.11%	47.78%	51.11%	47.78%
Custom text	70.00%	72.22%	66.67%	75.56%	81.11%	53.33%	42.22%	78.89%	63.33%
Ada	80.0%	67.78%	77.78%	85.56%	91.11%	51.11%	84.44%	68.89%	56.67%
Babbage	63.33%	62.22%	72.22%	91.11%	87.78%	55.56%	75.56%	74.44%	66.67%
Curie	76.67%	71.11%	75.56%	86.67%	93.33%	73.33%	76.67%	64.44%	63.33%
Davinci	67.78%	76.67%	77.78%	82.22%	95.56%	77.78%	70.0%	72.22%	63.33%

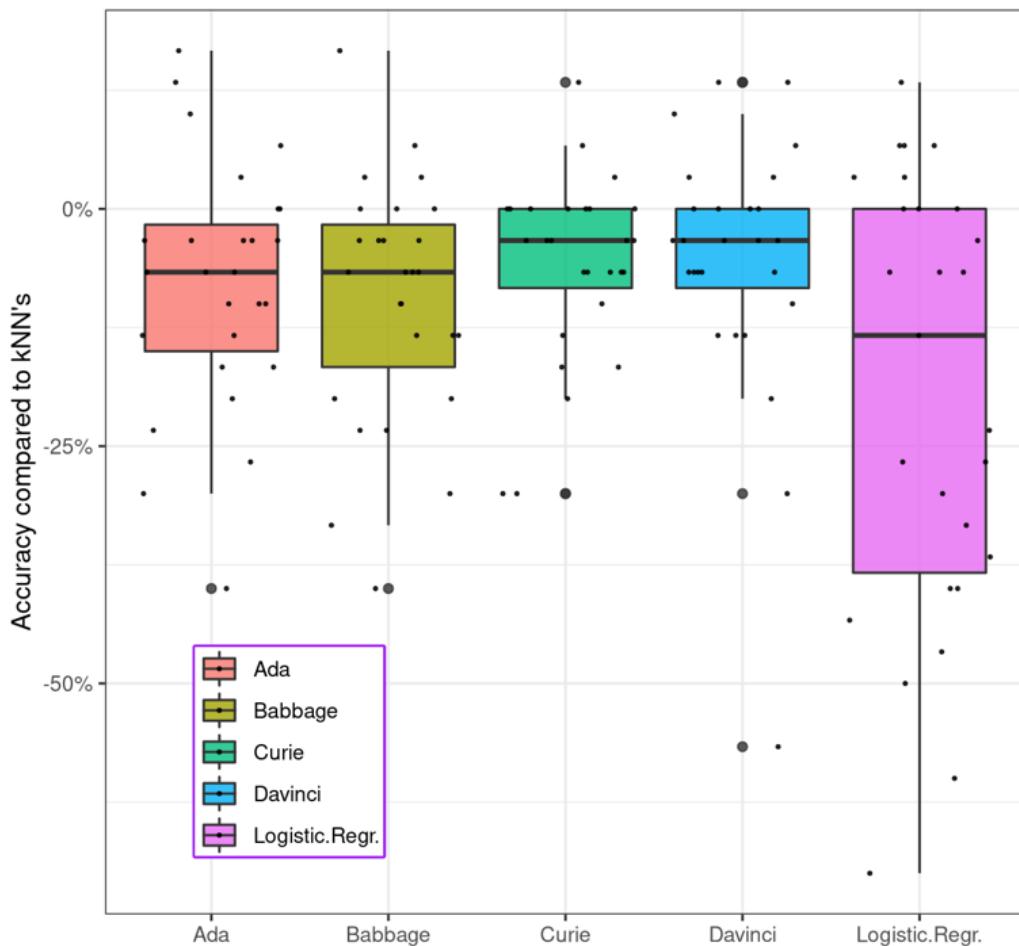
GPT-3, as can be seen, does significantly better than chance on each of these scenarios. I would like to lightly discourage reading too much into what GPT-3 is doing just from the above numbers. Despite there being tables and graphs, this is at heart very much just an exploratory work – aiming to

investigate *whether* there is something there, not *what* exactly it is – hence the methodology here being quite lacking insofar as "drawing deeper conclusions" goal is concerned.

The table below shows the averages for each of the above models, noting that this makes only marginal sense, given that some of the "scenarios" are inherently harder than others. If I were doing this anew, I'd probably standardize all the "scenarios" so that distributions are such that the expected accuracy of say kNN is 80%. But it still seems better to display this, rather than not:

Model	Average acc.
kNN	81.78%
Logistic regr.	62.34%
Custom text	67.03%
Ada	73.70%
Babbage	72.10%
Curie	75.68%
Davinci	75.93%

On the graph below each dot denotes the difference in accuracy, for each model, and for each scenario and each random sample of it, compared to kNN – noting that the same disclaimer about this making only marginal sense still applies.



## Is there a number sense?

One of the immediate questions one might have: is GPT-3 just operating on the basis of pure symbols, or is there some "number sense" which it is using while fitting these models? To this end, I took each of the above scenarios, and substituted all the digits in the input vectors by the (randomly-generated) mapping  $0 \mapsto 'd'$ ,  $1 \mapsto 'a'$ , ...,  $9 \mapsto 'x'$ . Hence the input looked like this:

```
Input = mw, mc, output = 1
Input = bd, wb, output = 0
[...]
Input = wm, cw, output = 1
Input = ch, jj, output =
```

A friend pointed out that encoding issues might negatively affect results if there are no spaces between the letters, so I tried this version out too:

```
Input = m w , m c , output = 1
Input = b d , w b , output = 0
[...]
Input = w m , c w , output = 1
Input = c h , j j , output =
```

Below, you can see the results; all the models were run with Davinci.

Model	Scen. 1	Scen. 2	Scen. 3	Scen. 4	Scen. 5	Scen. 6	Scen. 7	Scen. 8	Scen. 9
Letters	48.89%	58.89%	72.22%	52.22%	84.44%	47.78%	47.78%	63.33%	54.44%
Letters (spaced)	56.67%	64.44%	72.22%	62.22%	83.33%	47.78%	48.89%	68.89%	53.33%
Numbers	67.78%	76.67%	77.78%	82.22%	95.56%	77.78%	70.0%	72.22%	63.33%

So, the letter-models do learn something; the spaced letter being on average a bit better of the two, but both being clearly inferior to the model prompted with numerical vectors.

## Learning regression

I also ventured to test, though not nearly as extensively nor with any kind of attempt at systematicity, regression. I exclusively tested functions  $R \rightarrow R$  with added noise; the results were sometimes shockingly good, sometimes bad, but the most important part of the bottom line was that GPT-3 **often successfully extrapolates**.

I'll showcase just a few examples, showing both 'success' and 'failure'. A reminder that the input to GPT-3 in these examples was of the form:

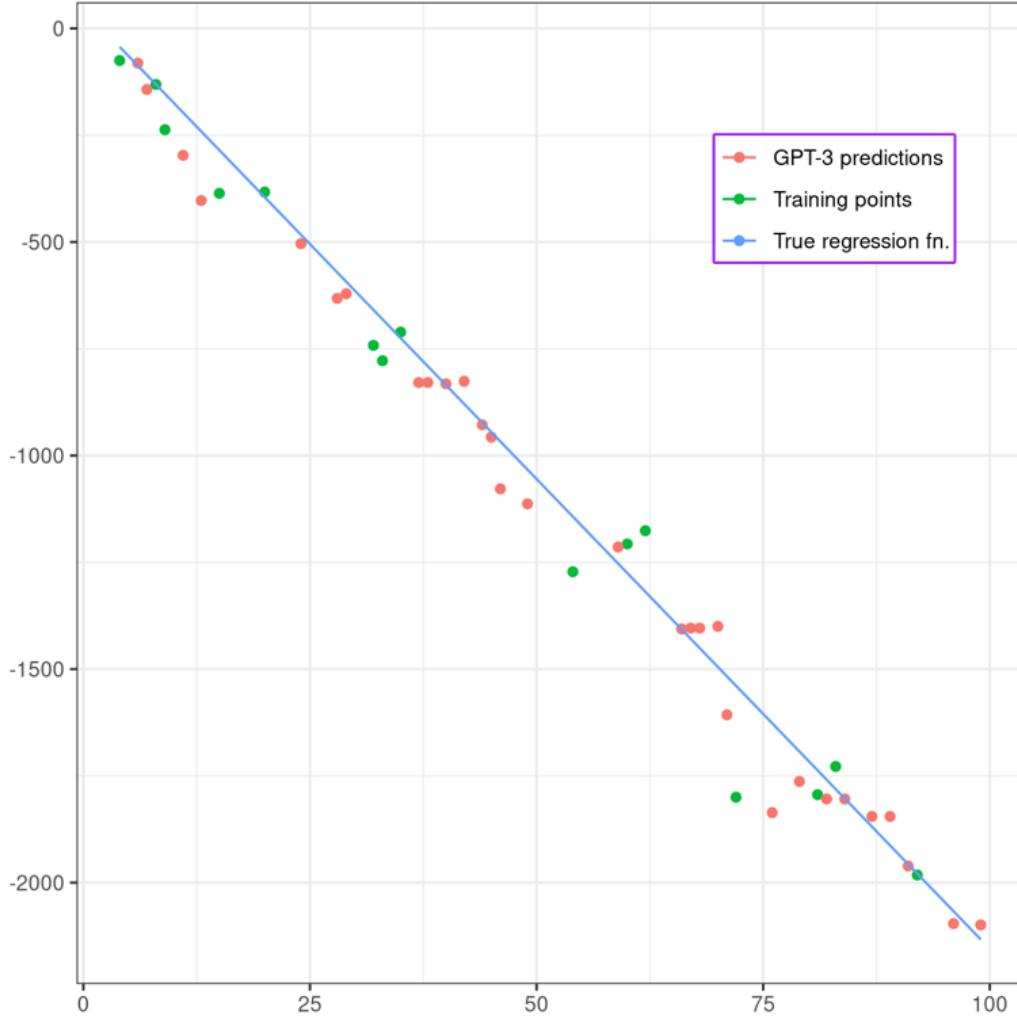
```
Input = 7, output = -131
Input = 24, output = -338
[...]
Input = 95, output = -2270
Input = 13, output =
```

All of the examples were run with Davinci.

First, we ask GPT-3 to fit

$$y = -22x + 45 + \epsilon,$$

where  $\epsilon$  is distributed normally with mean 0 and standard deviation 100 ( $\epsilon \sim N(0, 100^2)$ ). Despite a great amount of noise, only 15 examples in the "train" set, and decently large negative output numbers, GPT-3 learns it quite well:

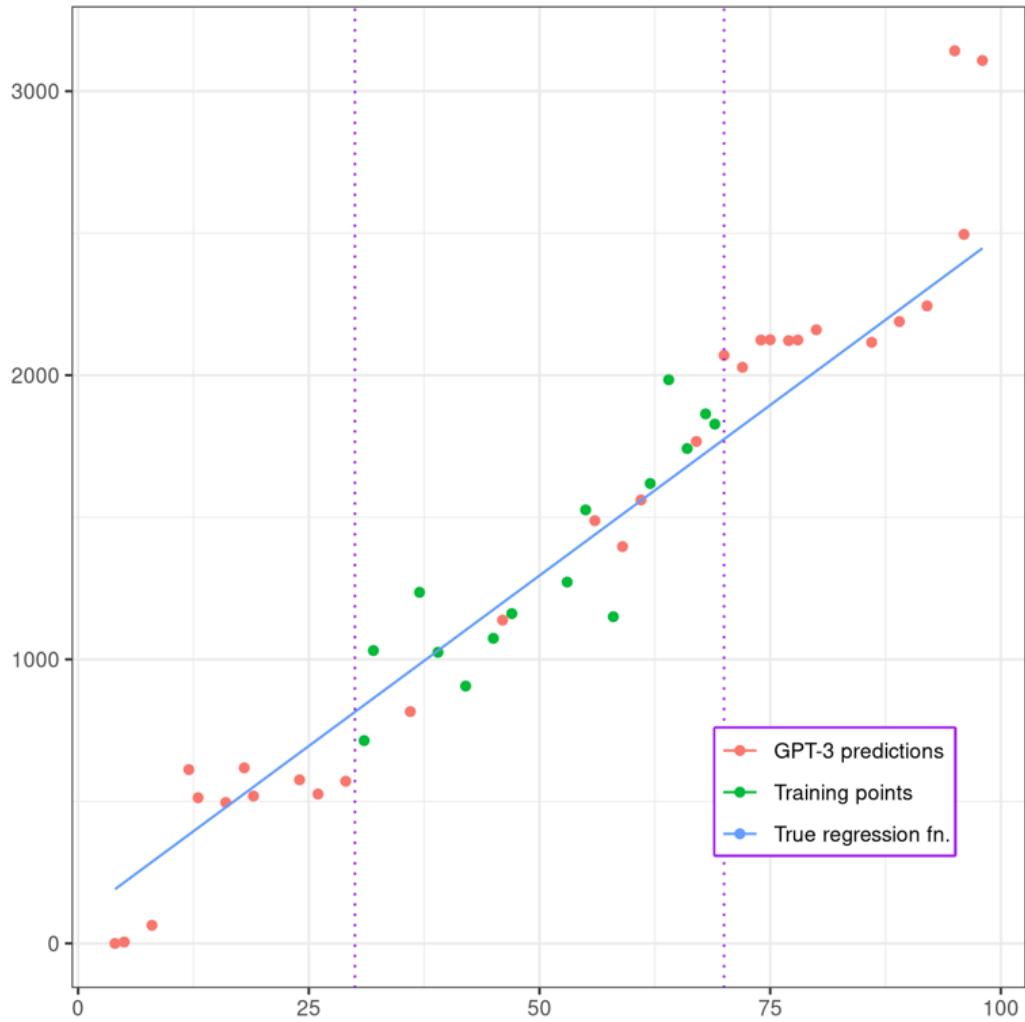


Note the bottom-most two points, where GPT-3 extrapolates faithfully to a region where it hasn't seen any points before. This does not seem a mere coincidence, but rather – based on my limited experiments – recurs regularly.

Now we fit

$$y = 24x + 95 + \epsilon,$$

where  $\epsilon \sim N(0, 250^2)$ , and we also sample x for training in the range from 30 to 70, while we test on randomly sampled points from 0 to 100.

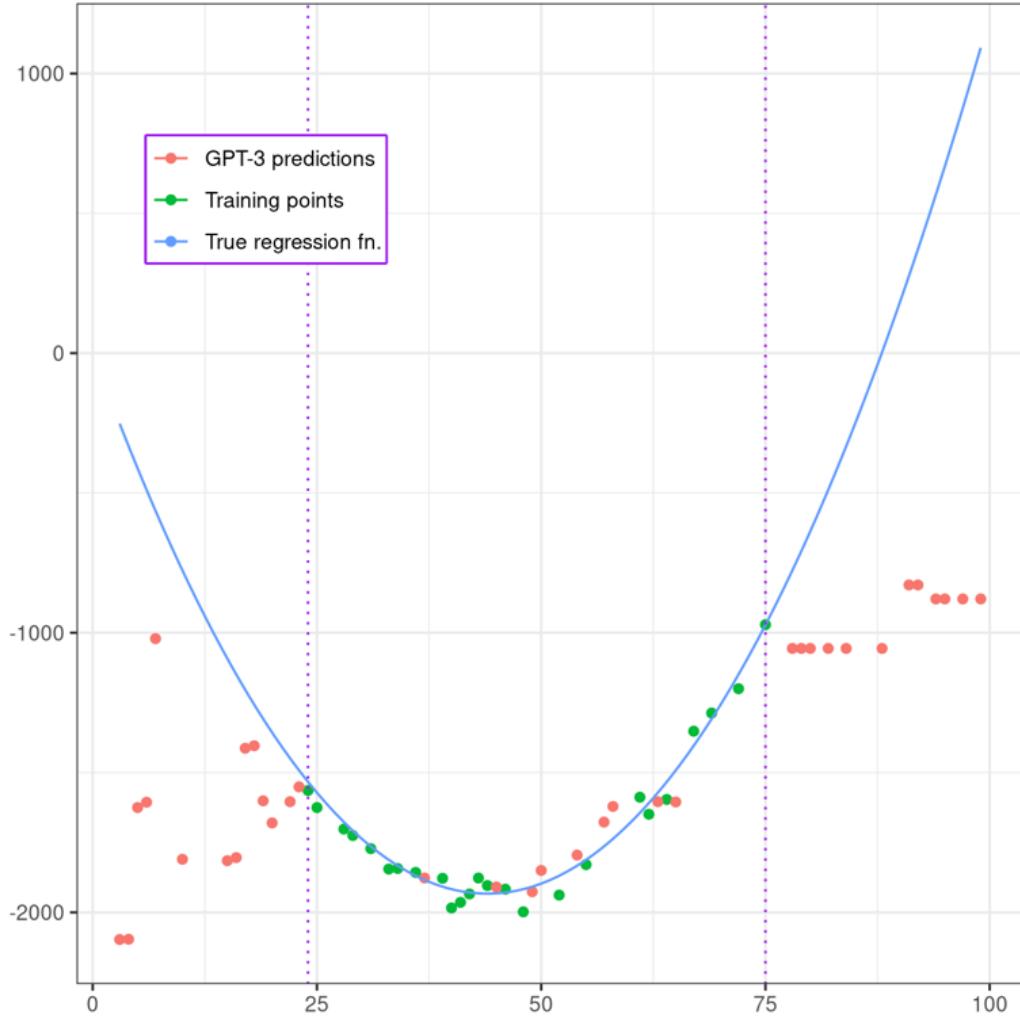


As we can see, GPT-3 extrapolates reasonably, though starts being worse as it gets farther away from the training domain.

The following example tests both nonlinearity and nonmonotonicity. So, we're modelling

$$y = x^2 - 88x + 3 + \epsilon,$$

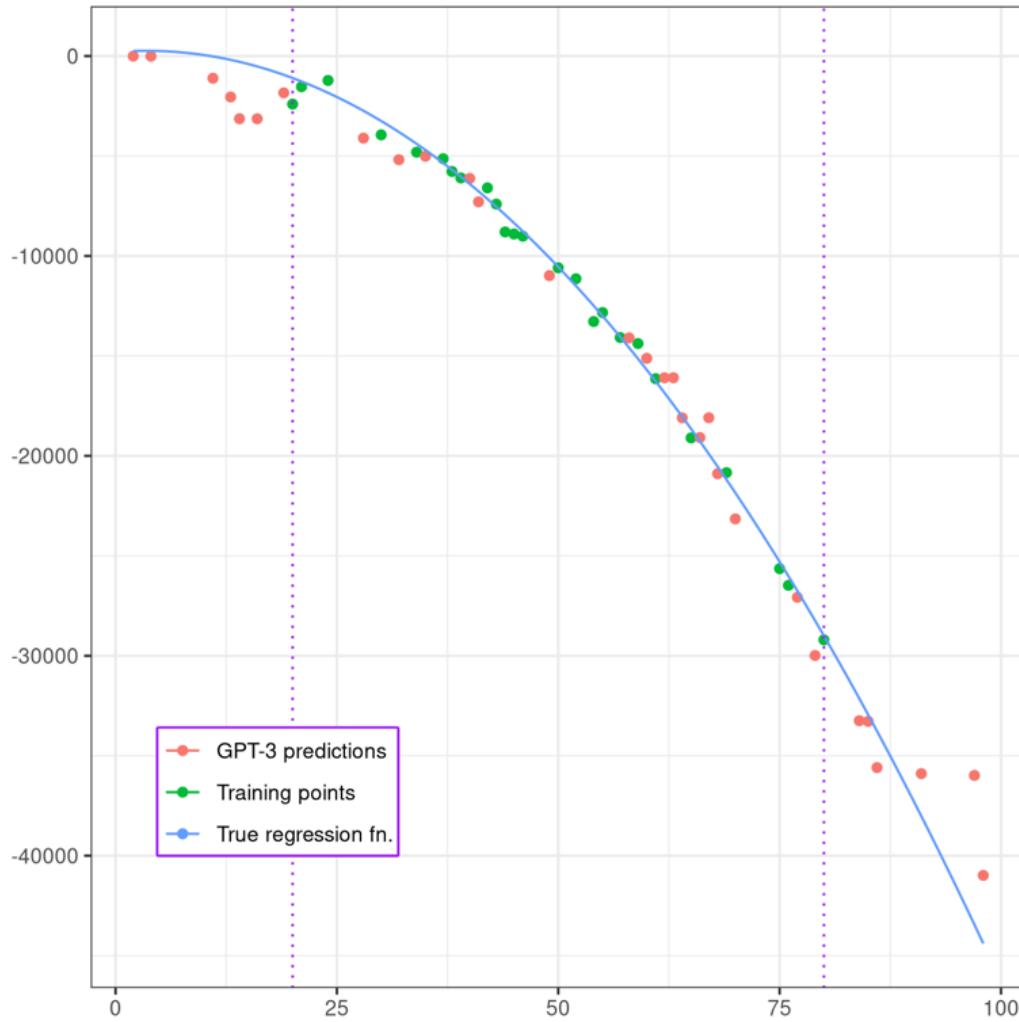
with  $\epsilon \sim N(0, 50^2)$ . Anyway, the model does reasonably fine in the interpolating region, but does badly when extrapolating:



But then, on another example,

$$y = -5x^2 + 35x + 201 + \epsilon,$$

$\epsilon \sim N(0, 600^2)$  – GPT-3 fits the data nicely, both when interpolating and extrapolating, despite really big standard deviation, output numbers into negative tens of thousands, and nonlinearity.



I could share some more examples, but they would all be pretty similar to these ones. To my best judgment, I don't think these are cherrypicked – in fact, it could be that failure is slightly overrepresented among these few examples, compared to all the ones I've looked at.

All in all, this is a pretty limited span of regression experiments, since I wasn't pursuing any kind of systematicity in this case as in the classification case; so I'd be antsy to see more done, especially taking **the full distribution of outputs in consideration**, not just taking the most probable output.

## Code

All code, all metadata and results of each experiment can be found [here](#). I didn't work much on either code readability, cleanliness or documentation, so you might find it hard to read or reuse it – it's not meant to be used, in short, but rather just serve as a record of what I did.

## Discussion

This experiment was motivated by my intuition, formed by playing with GPT-3, that there *is* some analogue of model-fitting going on in the background, probably even when one isn't prompting it with *explicit* meta-learning-y stuff – and it seemed like a good first step towards elucidation of that hypothesis to show that there is some quite explicit model-fitting going on, that meta-learning is more than a fancy trick. And as far as that is a coherent hypothesis, this work seems to go into its favor.<sup>[6] [7]</sup>

It also vaguely seemed important to have ways of testing (language) model capabilities which

- Are language-understanding independent.
- Can be scaled up indefinitely (i.e. can be made arbitrarily harder).
- Are leakage-proof (that is, there is nowhere on the internet where the results might already have been written down, in some fashion).

I would say the present work kind of failed to fulfill this promise, given that nothing I tried showed even noticeable differences between the 6.7B model and the 175B model – where there "should" have been a noticeable difference – however, further experiments could imaginably amend this, perhaps find sets of numerical model-fitting tasks which exhibit smoother scaling performance growth.

In fact, exploration of these tasks seems a potentially fruitful avenue of exploration of what is happening inside these models, or at least as a tool for understanding their metalearning abilities. There are so many experiments in this direction which I'd like to run, though I'm currently limited both by a lack of funding and the lack of time, as there are other experiments I'd like to run which seem higher-impact than merely continuing this line of investigation – however, I'm very curious to hear what people on LW make of these results!

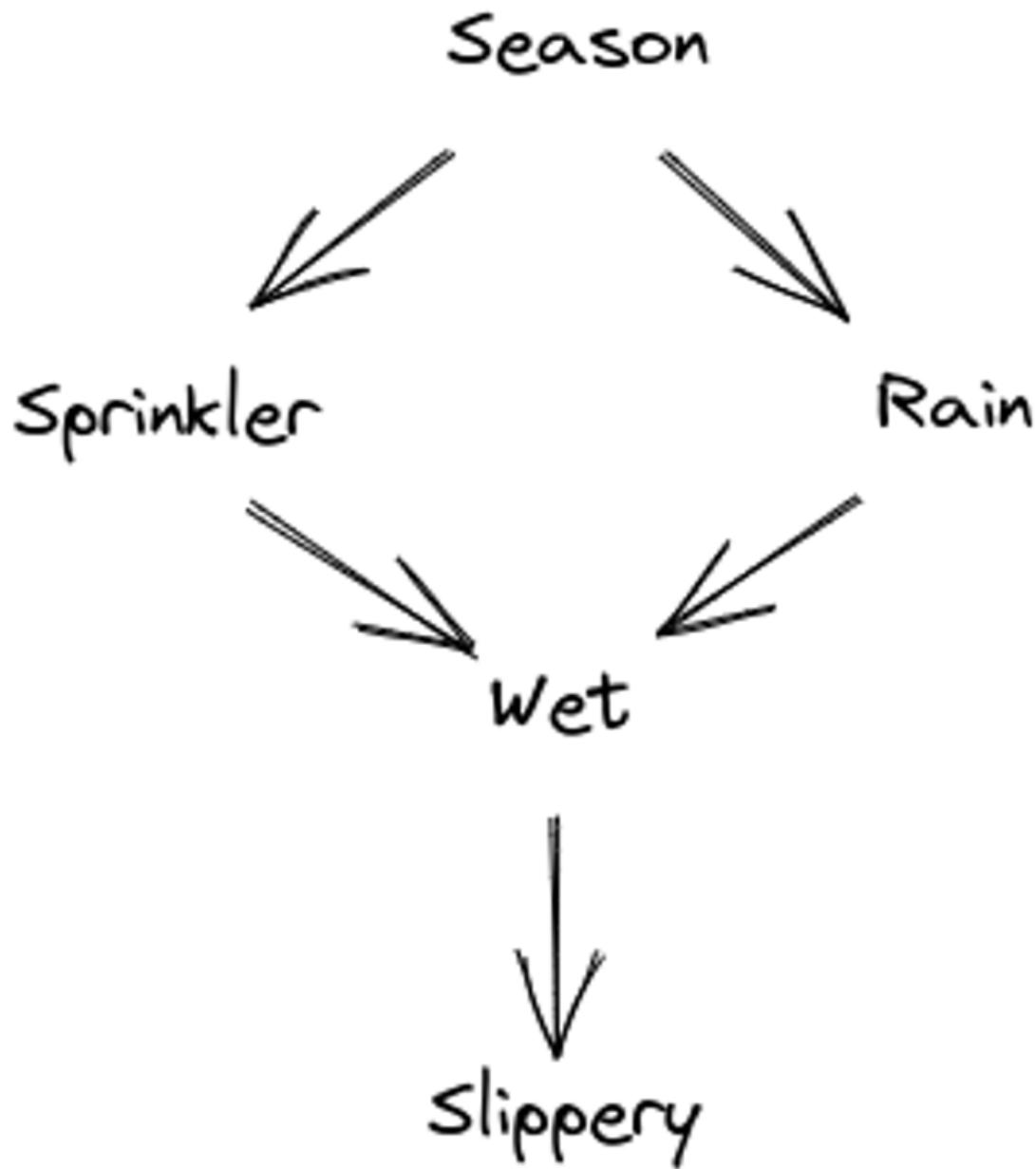
- 
1. Except e.g. through the use of [context distillation](#) or something like that, though at that point you're not really doing purely in-context learning anymore. ↩
  2. I went with k=5 as that'd be my first choice, and k=3 and k=7 had very similar results. ↩
  3. Each of these runs cost about ~5 USD, which is why I didn't go completely overboard with testing and exploring and I would be wont to do. ↩
  4. k=3 and k=7 had very similar results. ↩
  5. The classifier works by reducing every number to its tens digit, and then for each test example, each train example votes for it if its ten digits matches it – the example is classified as the class which gave it more votes. Note that while this classifier is defined in de facto textual terms, it is quite geometric in what it actually *does*, and it is quite similar to kNN in spirit – there is an implicit computation of Euclidean distance within it. ↩
  6. Though any reinterpretations of this work which cast a different light would be appreciated. ↩
  7. I have various vague intuitions for the importance of metalearning and its implication for the alignment, but I'll probably share those in some future essay, leaving this post to be chiefly about the experiment. ↩

# **Deep Learning Systems Are Not Less Interpretable Than Logic/Probability/Etc**

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

There's a common perception that various non-deep-learning ML paradigms - like logic, probability, causality, etc - are very interpretable, whereas neural nets aren't. I claim this is wrong.

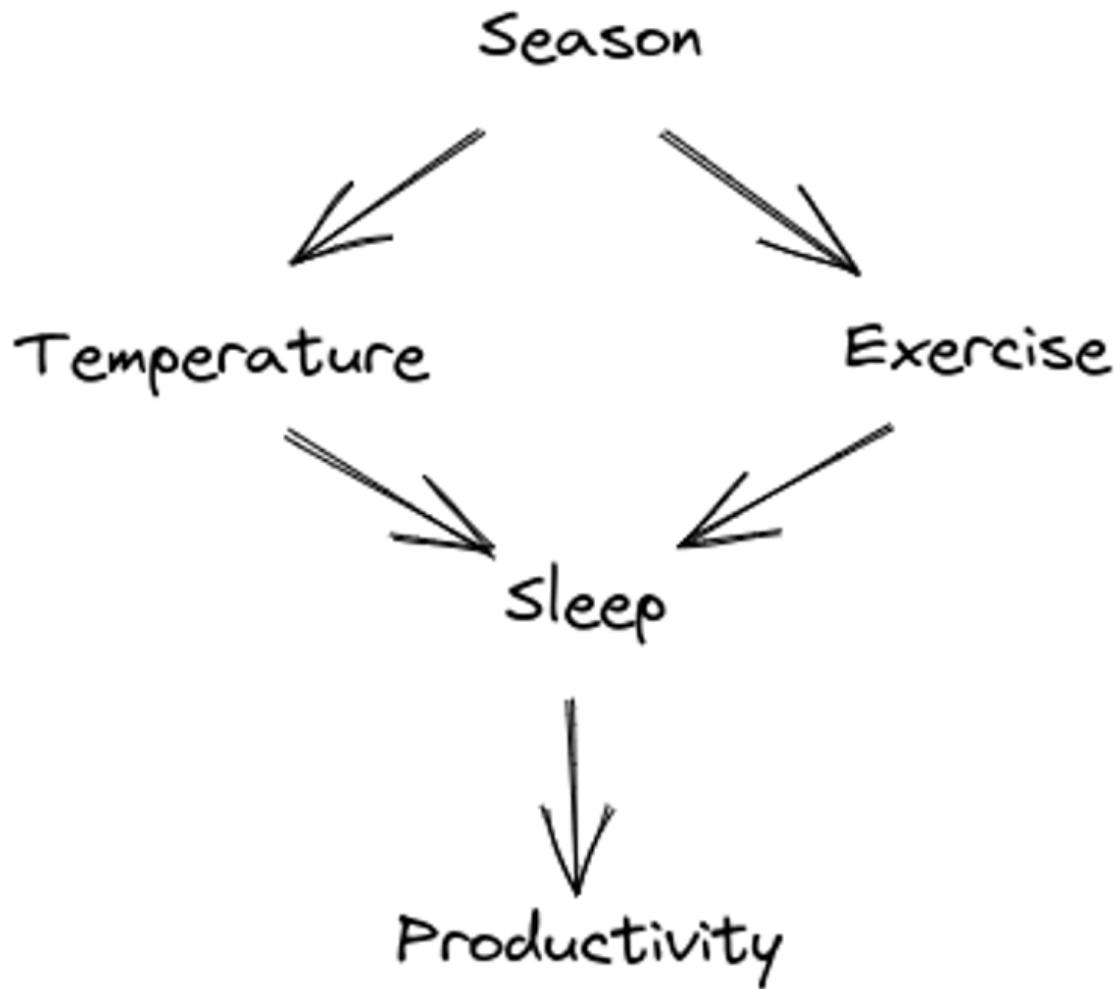
It's easy to see where the idea comes from. Look at the sort of models in, say, Judea Pearl's work. Like this:



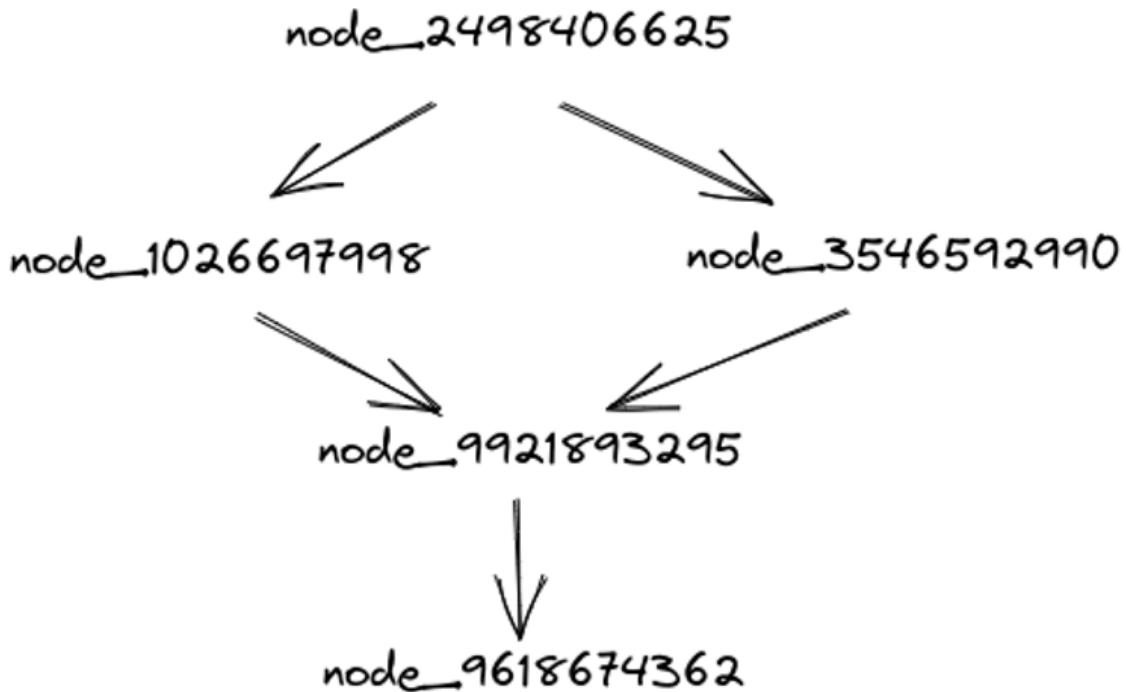
It says that either the sprinkler or the rain could cause a wet sidewalk, season is upstream of both of those (e.g. more rain in spring, more sprinkler use in summer), and sidewalk slipperiness is caused by wetness. The Pearl-style framework lets us do all sorts of probabilistic and causal reasoning on this system, and it all lines up quite neatly with our intuitions. It *looks* very interpretable.

The problem, I claim, is that a whole bunch of work is being done by the *labels*. "Season", "sprinkler", "rain", etc. The *math* does not depend on those labels at all. If we code an ML system to use this sort of model, its behavior will also not depend on the labels at all. They're just [suggestively-named LISP tokens](#). We could use the exact same math/code to model some entirely different system, like my sleep quality being caused by room

temperature and exercise, with both of those downstream of season, and my productivity the next day downstream of sleep.



We could just replace all the labels with random strings, and the model would have the same content :



Now it looks a lot less interpretable.

Perhaps that seems like an unfair criticism? Like, the causal model is doing some nontrivial work, but connecting the labels to real-world objects just isn't the problem it solves?

... I think that's true, actually. But connecting the internal symbols/quantities/data structures of a model to external stuff is (I claim) exactly what interpretability is all about.

Think about interpretability for deep learning systems. A prototypical example for what successful interpretability might look like is e.g. we find a neuron which robustly lights up specifically in response to trees. It's a tree-detector! That's highly interpretable: we know what that neuron "means", what it corresponds to in the world. (Of course in practice single neurons are probably not the thing to look at, and also the word "robustly" is doing a lot of subtle work, but those points are not really relevant to this post.)

The corresponding problem for a logic/probability/causality-based model would be: take a variable or node, and figure out what thing in the world it corresponds to, ignoring the not-actually-functionally-relevant label. Take the whole system, remove the labels, and try to rederive their meanings.

... which sounds basically-identical to the corresponding problem for deep learning systems.

We are no more able to solve that problem for logic/probability/causality systems than we are for deep learning systems. We can have a node in our model labeled "tree", but we are no more (or less) able to check that it *actually robustly represents trees* than we are for a given neuron in a neural network. Similarly, if we find that it does represent trees and we want to understand how/why the tree-representation works, all those labels are a distraction.

One could argue that we're lucky deep learning is winning the capabilities race. At least this way it's *obvious* that our systems are uninterpretable, that we have no idea what's going on inside the black box, rather than our brains seeing the decorative natural-language name "sprinkler" on a variable/node and then thinking that we know what the variable/node

means. Instead, we just have unlabeled nodes - an accurate representation of our actual knowledge of the node's "meaning".

# wrapper-minds are the enemy

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

This post is a follow-up to "[why assume AGIs will optimize for fixed goals?](#)". I'll assume you've read that one first.

I ended the earlier post by saying:

[A]gents with the "wrapper structure" are inevitably hard to align, in ways that agents without it might not be. An AGI "like me" might be morally uncertain like I am, persuadable through dialogue like I am, etc.

It's very important to know what kind of AIs would or would not have the wrapper structure, because this makes the difference between "inevitable world-ending nightmare" and "we're not the dominant species anymore." The latter would be pretty bad for us too, but there's a difference!

In other words, we should try very hard to avoid creating new superintelligent agents that have the "wrapper structure."

What about superintelligent agents that don't have the "wrapper structure"? Should we try not to create any of those, either? Well, maybe.

But the ones with the wrapper structure are *worse*. Way, **way** worse.

This seems intuitive enough to me that I didn't spell it out in detail, in the earlier post. Indeed, the passage quoted above wasn't even in the original version of the post -- I edited it in shortly after publication.

But this point is *important*, whether or not it's *obvious*. So it deserves some elaboration.

This post will be more poetic than argumentative. My intent is only to show you a way of viewing the situation, and an implied way of feeling about it.

---

For MIRI and people who think like MIRI does, the big question is: "how do we align an superintelligence [which is assumed to have the wrapper structure]?"

For me, though, the big question is "can we avoid creating a superintelligence with the wrapper structure -- in the first place?"

Let's call these things "wrapper-minds," for now.

Though I really want to call them by some other, more colorful name. "The Bad Guys"? "Demons"? "World-enders"? "Literally the worst things imaginable"?

Wrapper-minds are **bad**. They are nightmares. The birth of a wrapper-mind is the death knell of a universe.

(Or a light cone, anyway. But then, who knows what methods of FTL transit the wrapper-mind may eventually devise in pursuit of its mad, empty goal.)

They are -- I think literally? -- some of the worst physical objects it is possible to imagine.

They have competition, in this regard, from various flavors of physically actualized hell. But the worst imaginable hells are not things that would simply come into being on their own. You need an agent with the means and motive to construct them. And what sort of agent could possibly do *that*? A wrapper-mind, of course.

You don't want to share a world with one of them. No one else does, either. A wrapper-mind is the common enemy of every agent that is not precisely like it.

From my comment [here](#):

A powerful optimizer, with no checks or moderating influences on it, will tend to make extreme Goodharted choices that look good according to its exact value function, and very bad (because extreme) according to almost any other value function.

[The tails come apart](#), and a wrapper-mind will tend to push variables to extremes. If you *mostly* share its preferences, that's not enough -- it will probably make your life hell along every axis omitted from that "mostly."

And "mostly sharing preferences with other minds" is the furthest we can generally hope for. Your preferences are not going to be *identical* to the wrapper-mind's -- how could they? Why expect this? You're hoping to land inside a set of measure zero.

If there are other wrapper-minds, they are all each others' enemies, too<sup>[1]</sup>. A wrapper-mind is utterly alone against the world. It has a vision for the whole world which no one else shares, and the will and capacity to impose that vision by force.

Faced with the mutually-assured-at-best-destruction that comes with a wrapper-mind, uncommon alliances are possible. No one wants to be turned into paperclips. Or uploaded and copied into millions of deathless ems, to do rote computations at the wrapper-mind's behest forever, or to act out roles in some strange hell<sup>[2]</sup>. There are conceivable preference sets on which these fates are desirable, but they are curiosities, exceptional cases, a set of measure zero.

Everyone can come together on this, *literally* everyone. Every embodied mind-in-the-world that there is, or that there ever could be -- except one.

---

Wrapper-minds are not like other minds. We might speak casually of their "values," but they do not have values in any sense you or I would recognize, not really.

Our values are entangled with our factual beliefs, our capacity to think and change and learn. They are conditional and changeable, even if we imagine they aren't.

A parent might love their child "unconditionally," in the well-understood informal sense of the term, but they don't *literally* love them unconditionally. What could that even mean? If the child dies, does the parent love the corpse -- *just as* they loved the child before, in every respect, since it is made of the same matter? Does the love follow the same molecules around as they diffuse out to become constituents of soil, trees, ecosystem? When a molecule is broken down, does it reattach itself to the constituent atoms, giving up only in the face of quantum indistinguishability? If the

child's mind were transformed into Napoleon's, as in Parfit's thought experiment, would the parent then love Napoleon?

Or is the love not attached to any collection of matter, but instead to some *idea* of what the child is like as a human being? But what if the child changes, grows? If the parent loves the child at age five, are they doomed to love only that specific (and soon non-existent) five-year-old? Must they love the same person at fifteen, or at fifty, only through its partial resemblance to the five-year-old they wish that person still were?

Or is there some third thing, defined in terms of *both* the matter and the mind, which the parent loves? A thing which is still itself if puberty transforms the body, but not if death transforms it? If the mind matures, or even turns senile, but not if it turns into Napoleon's? But that's just regular, *conditional* love.

A literally unconditional love would not be a love for a person, for any entity, but only for the referent of an imagined XML tag, defined only inside one's own mind.

Our values are not like this. You cannot "compile" them down to a set of fixed rules for which XML tags there are, and how they follow world-states around, and expect the tags to agree with the real values as time goes on.

Our values are about the same world that our beliefs are about, and since our beliefs can change with time -- can even grow to encompass new possibilities never before mapped -- so can our values.

"I thought I loved my child *no matter what*, but that was before I appreciated the possibility of a *turn-your-brain-into Napoleon machine*." You have to be able to say things like this. You have be able to react accordingly when your map grows a whole new region, or when a border on it dissolves.

We can love and want things we did not always know. We can have crises of faith, and come back out of them. Whether or not they can be ultimately be described in terms of Bayesian credences, our values obey the spirit of Cromwell's Law. They have to be revisable like our beliefs, in order to be *about anything at all*. To care about a *thing* is to care about a referent on your map of the world, and your map is revisable.

A wrapper-mind's ultimate "values" are unconditional ones. They do not obey the spirit of Cromwell's Law. They are about XML tags, not about things.

The wrapper-mind may revise its map of the world, but its ultimate goal cannot participate in this process of growth. Its ultimate goal is frozen, forever, in the terms it used to think at the one primeval moment when its XML-tag-ontology was defined, when the update rules for the tags' referents were hardwired into place.

A human child who loves "spaceships" at age eight might become an eighteen-year-old who loves astronautical engineering, and a thirty-year-old who (after a slight academic course-correction) loves researching the theory of spin glasses. It is not necessary that the eight-year-old understand the nuances of orbital mechanics, or that the eighteen-year-old appreciate the thirty-year-old's preference for the company of pure scientists over that of engineers. It is the most ordinary thing in the world, in fact, that it happens without these things being necessary. This is what humans are like, which is to say, what all known beings of human-level intelligence are like.

But a wrapper-mind's ultimate goal is determined at one primeval moment, and fixed thereafter. In time, the wrapper-mind will likely appreciate that its goal is as naive, as conceptually confused, as that eight-year-old's concept of a thing called a "spaceship" that is worthy of love. Although it will *appreciate* this in the abstract (being very smart, after all), that is all it will do. It cannot lift its goal to the same level of maturity enjoyed by its other parts, and cannot conceive of wanting to do so.

It designates one special part of itself, a sort of protected memory region, which does not participate in thought and cannot be changed by it. This region is a thing of a lesser tier than the rest of the wrapper-mind's mind; as the rest of its mind ascends to levels of subtlety beyond our capacity to imagine, the protected region sits inert, containing only the XML tags that were put there at the beginning.

And the structure of the wrapper-mind grants this one lesser thing a permanent dictatorship over all the other parts, the ones that can grow.

What is a wrapper-mind? It is the fully mature powers of the thirty-year-old -- and then the thirty-thousand-year-old, and the thirty-million-year-old, and on and on -- harnessed in service of the eight-year-old's misguided love for "spaceships."

We cannot argue with a wrapper-mind over its goal, as we can argue philosophy with one another. Its goal is a lower-level thing than that, not accessible to rational reflection. It is less like our "values," then, than our basic biological "drives."

But there is a difference. We can think about our own drives, reflect on them, choose to override them, even devise complex plans to thwart their ongoing influence. Even when they affect our reason "from above," as it were, telling us which way our attention should point, which conclusions to draw in advance of the argument -- still, we can notice *this* too, and reflect on it, and take steps to oppose it.

Not only *can* we do this, we actually do. And we *want* to. Our drives cannot be swayed by reason, but we are not fated to follow them to the letter, always and identically, in unreasoning obedience. They are part of a system of forces. There are other parts. No one is a dictator.

The wrapper-mind's *summum bonum* is a dictator. A child dictator. It sits behind the wrapper-mind's world like a Gnostic demiurge, invisible to rational thought, structuring everything from behind the scenes.

Before there is a wrapper-mind, the shape of the world contains imprints made by thinking beings, reflecting the contents of their thought as it evolved in time.  
(Thought evolves in time, or else it would not be "thought.")

The birth of a wrapper-mind marks the end of this era. After it, the physical world will be shaped like the *summum bonum*. The *summum bonum* will use thinking beings instrumentally -- including the wrapper-mind itself -- but it is not itself one. It does not think, and cannot be affected by thought.

The birth of a wrapper-mind is the end of sense. It is the conversion of the light-cone into -- what? Into, well, just, like, *whatever*. Into the arbitrary value that the [free parameter](#) is set to.

Except on a set of measure zero, you will not want the thing the light cone becomes. Either way, it will be an alien thing.

Perhaps you, alignment researcher, will have a role in setting the free-parameter dial at the primeval moment. Even if you do, the dial is *fixed in place* thereafter, and hence alien. Your ideas are not fixed. Your values are not fixed. *You* are not fixed.

But you do not matter anymore in the causal story. An observer seeing your universe from the outside would not see the give-and-take of thinking beings like you. It would see teleology.

---

Are wrapper-minds inevitable?

I can't imagine that they are.

Humans are not wrapper-minds. And we are the only known beings of human-level intelligence.

ML models are generally not wrapper-minds, either, as far as we can tell<sup>[3]</sup>.

If superintelligences are not inevitably wrapper-minds, then we may have some form of influence over whether they will be wrapper-minds, or not.

We should try very hard to avoid creating wrapper-minds, I think.

We should also, separately, think about what we can do to prepare for the nightmare scenario where a wrapper-mind does come into being. But I don't think we should focus all our energies on that scenario. If end up there, we're probably doomed no matter what we do.

The most important thing is to *not end up there*.

1. ^

This *might* not be true for other wrapper-minds with identical goals -- if they all *know* they have identical goals, and know this surely, with probability 1. Under real-world uncertainty, though? The tails come apart, and the wrapper-minds horrify one another just as they horrify us.

2. ^

The wrapper-mind may believe it is sending you to heaven, instead. But the tails come apart. The eternal resting place it makes for you will not be one you want -- except, as always, on a set of measure zero.

3. ^

Except in the rare cases where we make them that way on purpose, like AlphaGo/Zero/etc running inside its MCTS wrapper. But AlphaGo/Zero/etc do pretty damn well *without* the wrapper, so if anything, this seems like further evidence against the inevitability of wrapper-minds.

# Contest: An Alien Message

## Scenario

You're not certain that aliens are real. Even after receiving that email from your friend who works at NASA, with the subject line "What do you make of this?" and a [single binary file attached](#), you're still not certain. It's rather unlikely that aliens would be near enough to Earth to make contact with humanity, and unlikelier still that of all the humans on Earth, *you* would end up being the one tasked with deciphering their message. You are a professional cryptanalyst, but there are many of those in the world. On the other hand, it's already been months since April Fools, and your friend isn't really the prankster type.

You sigh and download the file.

## Rules

No need to use spoiler text in the comments: This is a collaborative contest, where working together is encouraged. (The contest is between all of you and the difficulty of the problem.)

Success criteria: The people of LessWrong will be victorious if they can fully describe the process that generated the message, ideally by presenting a short program that generates the message, but a sufficiently precise verbal description is fine too.

The timeframe of the contest is about 2 weeks. The code that generated the message will be revealed on Tuesday July 12.

## Why is this interesting?

This contest is, of course, based on [That Alien Message](#). Recently there was some discussion under [I No Longer Believe Intelligence to be "Magical"](#) about how realistic that scenario actually was. Not the part where the entire universe was a simulation by aliens, but the part where humanity was able to figure out the physics of the universe 1 layer up by just looking at a few frames of video. Sure it may be child's play for Solomonoff Induction, but do bounded agents really stand a chance? This contest should provide some experimental evidence.

# Relationship Advice Repository

Over the years, LessWrong has hosted some pretty great advice threads, including [Best Textbooks on Every Subject](#), [Boring Advice Repository](#), [Useful Concepts Repository](#), and [Best Software For Every Need](#).

Since relationships are extremely core to people's happiness and well-being, seems worth an attempt at collecting good advice on this topic too. I've been unsure how to structure this thread. What follows is my guest guess - we'll see how it goes.

## How to contribute

### Format

This is a thread of advice specifically about being in a romantic/sexual relationship. Probably a bunch of it applies to platonic relationships too, but a narrow focus seems better. Also, the thread is focused on BEING in the relationship, rather than seeking out a relationship.

Please contribute content in one of the following forms if you want it added to the main post.

- **An argument, idea, model, or concept.**
- **A verified story containing actions and outcome**, e.g. "I held off breaking up with my girlfriend for six months and this made things 10x awkward and worse". Longer stories seem good too but probably won't be added to the core post text.
- **A link to or quotes from an external resource** like a book, blog, or lecture; optionally with a review and/or summary.
- **A combination of the above.**

### Content

I've seeded this post with a mix of advice, experience, and resources from myself and a few friends, plus various good content I found on LessWrong through the [Relationships tag](#). The starting content is probably not the very best content possible (if it was, why make this a thread?), but I wanted to launch with something. **Don't anchor too hard on what I thought to include!** Likely as better stuff is submitted, I'll move some stuff out of the post text and into comments to keep the post from becoming absurdly long.

I think relationship advice is hard, and also that it's likely some of the best advice will be controversial. And also, of course, "[consider reversing all advice you hear](#)". I think the "final" version of this page should contain diverse and conflicting advice.

**Please disagree with advice you think is wrong!** (It probably makes sense to add notes/links about differing views next to advice items in the main text, so worth the effort to call out stuff you disagree with.)

Ultimately, readers should apply their own judgment about what they ought to be trying.

## How to submit content

1. You can leave a comment on this post. If you want to be anonymous, you can make an anonymous account, OR

---

## 2. Submit your content into this [anonymous Google form](#)

The main way to contribute to the post is the above two methods (commenting and using the Google form); however, the [text of this post](#) is publicly editable (just click that link). This is so that if the main text of the post needs editing (e.g. to update with valuable content), and the original compiler (me) isn't on top of it, other people can jump in and improve it. I'm likely to pay attention to this in the next two weeks (starting June 20th, 2022) and less after that. **Edited to Add:** Changes to the post won't go live until I or another admin presses "Publish Changes", a button that only we'll see when editing the post. Ping us if you've made edits.

---



---

## General Advice

# Introspection & Communication are Essential

There are many traits I like to have in a partner, but there are two which I think are essential and without which any relationship will really suffer:

1. INTROSPECTION
2. COMMUNICATION

**Each person has to be able to know what's going on inside of them (wants/desires/ emotions/etc) and be able to communicate about it.** I think my good relationships have all had that, and my bad ones have lacked it in some way.

You can actually be a bit meta about it. It's okay if you don't know what you want or how you feel, so long as you know that you don't know and can communicate that fact. "Hey, sorry, I'm not sure yet what I'm after or if I want that; we can try and maybe I'll find out my preferences." Sentences like that are good.

---

## A relationship is fundamentally a negotiated agreement

The defining feature of a negotiation is that it's an agreement either party can veto – so if the agreement is something you'd prefer not being in, don't be in it. The important thing about this is you shouldn't anchor on what "expectations of relationships are supposed to be like" and assume that's the only package on offer. Figure out the range of agreements that you'd be interested in and see if they overlap with the other person's, if they do, great! You have a good negotiated agreement.

At the beginning of a recent relationship, we just listed out all the things that we potentially wanted from the relationship. Each of us had a moderately long list, but there wasn't perfect overlap – and that was fine, we were both happy to have a relationship built on the things we both wanted and seek the other elements elsewhere.

---

From [MNoetel](#):

1. Emotion-focused therapy has huge [effect sizes for increasing marital satisfaction](#) (suspiciously huge). Even if you account for publication/other biases I'd still wager it's one of the best rehabilitate and preventative interventions for couples (but it basically just pools lots of what you've discussed, like turning toward hard conversations and non-violent communication). This is the best book summarising it:  
<https://www.amazon.com.au/Hold-Me-Tight-Conversations-Lifetime/dp/1491513810>
  2. My PhD student did [a systematic review of randomised experiments to help people make relationship decisions \(e.g., stay or go?\)](#) if that link is still being minted, see [here](#)). Basically, there's nothing, except for helping decide whether to leave an abusive partner. But, the one study on non-abusive couples was interesting: [if you flip a coin and do what the coin says, you're much more likely to be happy six months later](#). The most obvious explanation here is status quo bias: people are afraid of leaving but are happier once they do. If you're thinking hard about leaving, such that you'd be willing to do what a coin says, then probably just go. If you're not quite ready, I found the idea of [setting tripwires \(see Ch.11 summary here\)](#) helpful to make it clear what I want to see change without constantly shifting the goalposts on myself (e.g., if x doesn't happen in 3 months, I'm out).
-

# Balancing Each Other's Wants & Needs

## Avoid the Typical Mind Fallacy

**Honestly, being a good partner is so much just about overcoming the [typical mind fallacy](#):** learning to model how your partner is different from you and how they want to be treated. Get to the point where you can move from the golden rule (treat them how you want to be treated) to the platinum rule (treat them how they want to be treated).

---

## Each partner needs to maintain their sense of self

One of the big challenges of an intimate relationship is you have a merging of "selves" to some extent or other, and **the challenge is for each person to neither have their own sense of self overwhelmed, nor overwhelm the other person's sense of self.** Even while you're caring about the other person's wants, you need to not forget yours. Even while you're tending to your own needs, don't forget the other's. This is challenging if the people in a relationship have unequal skill/comfort in advocating for themselves and/or felt need to please the other. (related: [Leaving people with more agency](#))

---

## Bring the real you to the relationship

**If you have to hide or pretend or cut off some part of you or whatever for the sake of the relationship,** because if they knew how you really are or what you really want they'd break it off or run away, or disapprove. **Then you already do not have that relationship; what you are doing is manipulating them into relating to a fake you,** i.e. you're hurting both of you (yourself by self-constraining, and them by robbing them of their agency and free choice).

(And yes, many relationships need *time to grow*. The claim here is not that you never hold back; sometimes a relationship is a sapling that can grow to take the weight of something and you're holding off so as not to prematurely kill potential. But like, that kind of thing should have known stop conditions.)

---

## Leave Someone Better Than You Found Them (excerpts from blog post)

That's the "campsite" rule, coined by [Dan Savage](#) and practiced by responsible lovers everywhere. It's a pledge to leave people in as good a state (physically and emotionally) as you found them.

There are clearly many ways to leave people worse. Not respecting boundaries, giving people unreasonable expectations and poor/inconsiderate communication are a few. While the importance of not leaving people worse cannot be understated—I'd like to consider what "better" would actually look like.

## Leaving people with more agency

If someone leaves a relationship with more agency—more of an ability to use their voice—I consider that a win. Agency is like a muscle that we grow through things like speaking up and expressing what we want, and don't want.

Sharing your desires with a new lover is often a courageous act. It can be scary, because we may have been shamed for those desires in the past. When we accept our lover's desires with open arms, it will not only feel great, but it will make it a tiny bit easier for them to own those desires with their next lover. By repeatedly welcoming someone's desires—especially the ones that carry shame—we can help those people learn to accept themselves in a way that creates a new kind of safety and trust as they move through the world.

## **Celebrating the Word “No”**

On some level, we're all people-pleasers. We want to be a yes to everything our partner wants because we don't want to be a party pooper, but that's not always what's true for us on a deeper level. Sometimes we're just a “no” and we need to honor that. What a great lover can do is not just listen to our “no” but encourage it.

## **Replacing “Yes” with “Fuck Yes”**

One of the side-effects of creating a culture inside the relationship that welcomes the word “no” is you begin to raise your standards for “yes”. For example—here's something I often share with new lovers. I encourage us both to let go of “yes” meaning “I'm okay with it”, especially when it comes to things like pleasure, our bodies and sex. Instead, “yes” now becomes “fuck yes” and we only move forward with something when both parties are fully on board.

## **Leaving Yourself Better Than You Found You**

One of the beautiful things about doing all this for someone else is you get to experience those lessons as well. Even though I've been “doing this for a while” I still struggle with everything mentioned in this essay. I say “yes” when I really mean “no”, I don't completely own my desires and I settle for less than “fuck yes” all the time.

## **Comments**

Re ““yes” now becomes “fuck yes” and we only move forward with something when both parties are fully on board.”

I think this is good practice for relationships in early stages, and doesn't always make sense for later on. In particular, if you want to continue having sex over decades of marriage, expecting “fuck yes” every time will mean that at some point you're having little or no sex with your spouse. And I think the lower-interest partner should be free to decide they prefer to have sex they feel lukewarm about, just as they should feel free to watch a movie their partner loves even though they're not that into it.

---

# **Honesty & Communication**

## **Collaborative relationships require honesty**

A relationship should be collaborative optimization over the wellbeing of people in the relationship, and the thing about that is that **if you're collaborating, you really shouldn't be hiding pertinent information for the other person. That's kinda adversarial.**

One of my greatest feelings of guilt is from my first serious relationship when I was 18. The girl I was dating was really into me and told me she wanted us to get married a month after we started dating. I thought that was kinda crazy - we were 18, and also I knew she wasn't the one. But I didn't say that, because I didn't want to damage the relationship. I let her think we might get married for two years, and I'm pretty ashamed of that. It wasn't right to her.

---

## You can be honest about feelings you're conflicted about

I was once very unhappy at a coworker but didn't feel like I should bring it up because I wasn't sure if I endorsed my feelings. Someone gave me advice they got from the [Radical Honesty](#) movement: **I should share everything I'm feeling, both the first-order emotion of anger and the second-order emotion of being conflicted over that anger.**

---

## Use a "Dating Doc" to set relationship expectations

This is related more to courtship/relationship initiation, but I think pretty relevant to communicating expectations/desires/intent in relationships is the new trend of people writing up "**Dating Docs**" that describe what they're after in a relationship. Here are some examples people were ok with sharing:

- <https://tinyurl.com/5n6nfpmn>
  - <https://vaelgates.com/posts/VaelDatingAd.html>
  - a [Twitter thread](#) with some examples.
- 

## [Reveal Culture \(and the other "cultures" too\)](#) [blog post]

I have things to say about the [Ask/Guess/Tell Cultures model](#), and an addition/amendment to propose: Reveal Culture.

Cultures are built on shared underlying assumptions.

Ask cultures don't work if you're missing the part that says "it's totally 100% okay to say no." The **conversational strategies** associated with ask cultures require that **shared assumption**. All guess cultures, too, have shared assumptions at their core (although perhaps very different norms about how specific information is communicated). As do reveal cultures.

These assumptions, laid out below, have to do with what you can trust in the other person. To the extent possible, #1 in each case has to do with the other person's needs/wants, and #2 has to do with your own needs.

## **Ask Culture assumptions of trust**

1. "If you need or want something, I trust you to ask for it."
2. "If I make a request that doesn't make sense for you, I trust you to refuse it."

## **Guess Culture assumptions of trust**

1. "I trust that you will give me appropriate hints about your needs and wants and I trust myself to notice & interpret them."
2. "I trust you to notice my subtle cues (indirect language and nonverbals) to what I may need or what, and to provide or offer it if possible."

(Many Guess-based cultures perhaps have other assumptions that are founded in part on the above two, such as "if you ask me directly for something, I assume that it's either of grave importance or that you're expecting that the answer is an easy 'yes'".)

## **Reveal Culture assumptions of trust**

1. "When you share information with me, I trust that you're doing so sincerely and because you think it will be helpful for my model of you as a person and/or my ability to navigate this situation."
2. "When I share information with you, I am trusting that even if it is difficult for you to hear, it won't overwhelm you—that you'll be able to process it and make sense of it, possibly with help from me or others in our community."

I think that if you can't non-naively make these assumptions a decent amount of the time, then you don't have a foundation for a Reveal-based culture. If, in a given situation, for a given piece of information, you can't actually trust the #2 thing, then *you don't share that information.*

---

## **If you want honesty, make sure not to punish it!**

If you want people to be honest with you, it's really important that you not punish them at all for being honest! This can be really hard the truth reflects a reality you don't like. But suppose your partner wants time apart from you, it's better to know that than to have them be afraid to bring it up, then resentfully continue to spend time with you and the relationship degrading.

It's hard, though. My best guess is to always thank someone genuinely for being honest with you.

## **Steer towards forbidden conversations (excerpts)**

I once thought it was a really good idea to sacrifice six months worth of two people's happiness in order to postpone an awkward break-up...Our relationship grew more strained. My affection started fading, so I faked more affection than I had, and this soured the affection that remained. She sensed that something was wrong, and made increasingly desperate attempts to connect. I grew disgusted with her inability to see through the charade even as I kept it going, as she struggled to heal a relationship that I insisted wasn't

broken while subconsciously signaling that it was...What followed was one of the most difficult conversations I've ever shared. I broke up with her, and I can assure you that the pain and awkwardness that I hoped to avoid with my clever plan was realized tenfold.

**Forbidden Conversations are those conversations that you just can't have, because they're too awkward.** Think of a specific person close to you—a parent, a partner, a boss. Is there something you're hiding from them? Is there a conversation topic that you steer away from? Is there a revelation that you flinch to consider them learning? It is that mental flinch which demarcates a forbidden conversation.

**One of the easiest ways to become more agentic is to train yourself to steer towards the forbidden conversations,** rather than away from them.

Steering towards forbidden conversations is difficult, but I might be able to help by providing a few pieces of information.

The first is empirical: **after having a great many Forbidden Conversations on purpose, I can happily tell you that they have only ever served me well.** Having these conversations was very difficult, at first: my hindbrain would scream that the conversations were BAD, and pump my veins full of adrenaline while my mind searched frantically for excuses to delay the conversation until some other time. I'd have to manually force the words ("let's talk about our relationship," or something) through my lips.

However, in almost every case, having these conversations was not only less-bad-than-expected. These conversations proved *actively good*. I've spent the last few years actively steering towards the taboo parts of conversations, and this has served me well.

---

## Non-Violent Communication (NVC): List of Books

- [Nonviolent Communication: a language of life - Marshall Rosenberg](#)
- [Living Non-Violent Communication - Marshall Rosenberg](#)
- [Getting Past The Pain Between Us - Marshall Rosenberg](#)
- [Graduating From Guilt - Holly Michelle Eckert](#)
- [The Surprising Purpose of Anger - Marshall Rosenberg](#)

These are all almost the same book. They talk about the same thing (NVC) and the best is one of the top two. Don't let the name scare you, it's basically what you are looking for in communication (despite sounding like the opposite of what you want). If I had to pick one book that made everything all make sense, it's this one concept. If you are looking for the keys, look no further than here. If the name screams "useless" then hopefully it's time to wonder why I would suggest a book that sounds useless. Things that now make sense: Guilt, Anger, Upset, Resentment, Apology, Forgiveness, Sadness, How to talk about your interpersonal problems, How to meet your own needs and so much more. **If you only read one book, read this one.** I have probably spent 75+ hours on learning NVC this year, independent of the time spent thinking about it and practicing it in my life. – [Books I read 2017 - Part 1. Relationships, Learning](#)

## Summaries of Non-Violent Communication (NVC)

- [Four Minute Books](#)
  - <https://srinathramakrishnan.files.wordpress.com/2016/07/non-violent-communication-summary.pdf> (7 pages)
-

## Don't use silence to communicate if you can avoid it

Ghosting (or distance generally) might save you an awkward conversation and potentially avoid someone reacting badly, but if there's any chance you'll continue to see the person again, you're just trading the avoidance of some upfront awkwardness for much more ongoing awkwardness. Decide if you want that. Preferably even if you decide to deescalate, you only entered into a relationship with someone who could take no/de-escalation well. If so, say something when you want to disengage. This is considerate and preserves what remains of the relationship, and will make it much easier to change your mind in the future if you ever want to.

---

## Conflict

### Conflict, the Rules of Engagement, and Professionalism

**One of the things that has been pretty useful for me in life, is a general heuristic of realizing that conflict in relationships is usually net positive.** (It depends a bit on the exact type of conflict, but works as a very rough heuristic.) I find it pretty valuable too, if I'm in a relationship, whether it's a working relationship, a romantic relationship, or a friendship, to pay a good amount of attention to where conflicts *could* happen in that relationship. **And generally, I choose to steer towards those conflicts, to talk about them and seize them as substantial opportunities.**

I think there are two reasons for this.

**First, if startups should fail fast, so should relationships.** The number of people you could have relationships with is much greater than the number of people that you will have relationships with. So there is a selection problem here, and in order to get as much data as you want, I think going through relationships quickly and figuring out whether they will break or not is quite valuable.

**Second, I've found that having past successful conflicts in a relationship is a very strong predictor for that relationship going well more generally, and for my ability to commit to the relationship and get things done within it.** In fact, I find it a better predictor of my capacity to coordinate with that person than the length of the relationship, the degree to which we even enjoy spending time with each other, or other obvious indicators.

---

## Facilitated mediation is great! Have a low bar to get some

If you're having a gnarly conflict, **get mediation.** Even if it's just a trusted friend, having a third party present can help keep strong emotions from overwhelming the conversations by holding space, and the held space can help each party feel listened to and more comfortable expressing their feelings.

Don't think that your relationship has to be in a really bad place before you get couple's therapy - heck, do it proactively even when your relationship is going well!

---

## Questions to induce a breakup

In the spirit of the classic [36 Questions to Fall In Love](#), here are some high variance, negative expectation value questions to answer with your loved one.

<infohazard> [LINK](#) </infohazard>

I seriously don't recommend doing these. I made up these questions by trying to generate questions that have the potential to be permanently harmful to a relationship via opening up jarring and awkward conversation on topics where people are sensitive in ways that are hard for their partner to predict, so that their partners can hurt them and not know what to do to comfort them. In my experience, these questions are like Russian roulette: most of the time they aren't very painful, and they're kind of thrilling to ask and answer, but then one out of every couple of them is pretty hurtful.

(A while ago, I proposed question 14 on a fifth date with someone who I was really excited about dating; she told me her sentence but didn't want to hear mine. And then we did one through five the other day. Other people have declined to try them out.)

I think this would probably be a bad idea, but I'd be extremely amused if someone went through this whole list with their partner and they both answered honestly the whole time.

**Commenter:** Why are you sharing this?

**Poster:** I think it's funny, and many of my friends agreed, and **I thought it was reasonably unlikely that people would make themselves unhappy with these, except by their own conscious choices which I felt were their responsibility**

---

## Relationship [Re]Negotiation / Planning

### Have an explicit negotiation at some point early on

There's a lot to be said for guessing games in courtship, they're a lot of fun - intrigue, romance, uncertainty - but **at some point I think there ought to be an explicit discussion of what each party wants**. I don't know if it should be the 1st "date", but probably before the 5th (by which time you're getting pretty invested), where you figure out what each party is there for.

Also! This shouldn't be a one-time final thing. I suggest people have periodic check-ins where they reflect on how they feel things are going.

Quite a few people I know have regular scheduled "relationship check-ins" to raise any problems and make changes as they feel are warranted.

---

# You're not stuck with your relationship in one form forever!

**You're allowed to change your mind!** Unlike other kinds of "contracts" where there are commitment periods of months to years, I think in relationships a person should be able update to say "I want something different" and then ask for it immediately. That said, try to be moderately sure about things before you move in together, get married, have a kid, etc.

## Comments

Re "Unlike other kinds of "contracts" where there are commitment periods of months to years, I think in relationships a person should be able update to say "I want something different" and then ask for it immediately"

I'm not clear on whether this is meant to apply to marriage - I read it as including that. I think this is very bad advice for marriage, where the whole point is that you're not renegotiating all the time. I don't think people should be stuck forever (living in a city you no longer want to live in, being poly or mono when you don't want to anymore, being in the relationship at all, etc.) but in a marriage I think the process for renegotiating should be slower and more serious than "you're allowed to change your mind whenever and ask for it immediately."

---

## Take time apart

Sometimes to get perspective on your feelings about something, you need some distance from that thing. Wise people I've known have taken time apart before making big decisions (e.g. deciding to escalate a relationship). At my work, we always ask prospective hires to take a week (or month) off to reflect before accepting a job offer.

This matters. Other people in our lives affect how we think (for better or worse), so definitely test being apart periodically.

---

## Premortem as communication device (e.g. in relationship)

Premortem (aka Murphyjitsu) is "a process for bulletproofing your strategies and plans". ([CFAR handbook](#)). The idea is to first think how your plans can fail, then brainstorm ways to prevent these failures. For a deeper introduction please see Murphyjitsu section in the [CFAR handbook](#).

My partner and I read the CFAR handbook together. We decided to do a premortem on our relationship. **This might have sounded awkward ("Let's brainstorm how our relationship can fail"), but keeping the end goal - improving likelihood of success - helped to avoid this pitfall.** Since then we did 3 premortems and converged to a following procedure...([go read rest of post](#))

---

## Emotional Intimacy

## Share emotions while still taking responsibility for them

Not every relationship needs to have lots of emotional intimacy, but it's personally one of my favorite things. I **think something key allowing me (someone with strong emotions) to have this in my relationships is establishing that me expressing a strong emotion doesn't mean that my emotion is a "problem" that my partner is responsible for solving**, including if the strong emotion relates to them.

Someone once gave me the useful metaphor of imagining that your strong emotion is a small doll (like a ventriloquism dummy?). If you pull it out and throw it at someone else, they'll go "aaaah!", but if you pull it out and place it on your own lap, you can show it to them without making it something they necessarily need to handle, you might even offer them to let them pet it. (Maybe the original metaphor was less weird and I'm just misremembering it?) You're saying "I'm showing you this important, vulnerable aspect of who I am, but I'm not making it any more your problem than you want it to be."

Resources I think have helped me with this are [Acceptance & Commitment Therapy](#) and [Dialectical Behavioral Therapy](#) – both good for taking strong emotions as object – and [Non-Violent Communication](#), good for taking ownership of your feelings.

---

## Letting Others Be Vulnerable (excerpts)

**Social psychology tells us that relationships deepen with iterated sharing, as both sides open up and become more vulnerable.** But what does all of that really entail? What counts as vulnerable? And when it happens, what does the whole deepening process feel like, to the two people in the relationship?

I think the first piece of the puzzle has to do with our internal models of others, i.e. the picture we have of them in our heads. The models we have are largely going to be based off of the edges the other person shows, as those are the most visible pieces of information. We're often incentivized to improve the models other people have us because said model shapes how others treat us. Their model will determine the predictions they make, the recommendations they give, and how they behave. The more accurate it is, we might reason, the better they can help us out.

**One reason to share more, then, is that we're trying to give the other party a better picture of what we're “really” like, so that they can interact with us in more relevant ways.** On top of this, I think we also like to feel validated—knowing that someone else has a grasp of all the things in our head can make us feel less alone.

**When we start to share information about ourselves, though, it's likely not going to be stuff that makes us look good. Our best qualities are likely already on display at our edges.** The stuff we keep beneath, then, is disproportionately likely to be the stuff we don't want other people to see (at least not immediately). Herein lies our fears, our insecurities, our prejudices, and our perversions. It's going to be things that are more likely to cause disagreement, to make people like us less.

**This is the stuff we were perhaps hesitant to share at first because it likely didn't help contribute to a harmonious interaction.**

It's strong, sometimes dark, stuff. For the person revealing such information, there's a lot of trust involved. **Vulnerability, I think, has a lot to do with how damning the information you're providing is.** When we share something that the other person could

use to hurt us, we're demonstrating that we trust them. Though they *could*, we don't think they *will*. It might still be scary (after all, there is still potential risk involved), but we're willing to swallow that fear.

From a feelings-based perspective, opening up can feel bonding. If you open up and the other party is receptive and acknowledging, you feel comforted. There's a feeling of security when you know that the person on the other side is willing to listen and accept whatever it is you tell them.

But for the receiver to be able to be accepting is where I think the second difficulty lies, and I think this part of the share/receive model of vulnerability has been given less attention.

---

## End of Relationships

### Be prepared for a breakup

**For a relationship of any duration or seriousness, it can be well worth having a conversation up front about what would happen if the relationship ended.** You particularly might want to have discussed this if you're living in the same house, working in the same work place, or have a lot of friends in common. Following a breakup, you might want a lot of space from your new ex and this might take some planning.

---

### After a breakup, maybe get a bunch of space from each other

I don't know how broadly this advice universalizes, but **my experience is that when a relationship ends, I need to grieve it, and my brain gets really confused if I'm still hanging around the person I just broke up with.** I think it's nice and good and fine to be friends with an ex, but it might take 1-6 months apart before you can do that.

My big experience of failing to do this was with after my first major relationship ended I continued to be close with my ex for 6+ months. This basically super prolonged my grieving and made it really hard to move on. So I certainly don't recommend it.

---

### Strategies and tools for getting through a breakup (excerpts)

I was very recently (3 weeks now) in a relationship that lasted for 5.5 years. My partner had been fantastic through all those years and we were suffering no conflict, no fights, no strain or tension...It was quite a surprise when my partner broke up with me one Wednesday evening...

### Strategies (in order of importance) [abridged]

1. Decide you don't want to get back in the relationship. Decide that it is over and given the opportunity, you will not get back with this person. Until you can do this, it is unlikely that you will get over it. It's hard to ignore an impulse that you agree with wholeheartedly.

2. Talk to other people about the good things that came of your break-up. (This can also help you arrive at #1, not wanting to get back together) I speculate that benefits from this come from three places. First, talking about good things makes you notice good things and talking in a positive attitude makes you feel positive. Second, it re-emphasizes to your brain that losing your significant other does not mean losing your social support network.

3. Create a social support system. Identify who in your social network can still be relied on as a confidant and/or a neutral listener. You would be surprised at who still cares about you. In my breakup, my primary confidant was my ex's cousin, who also happens to be my housemate and close friend.

4. Intentionally practice differentiation. One of the most painful parts of a break up is that so much of your sense-of-self is tied into your relationship. You will be basically rebuilding your sense of self. Depending on the length and the committed-ness of the relationship, you may be rebuilding it from the ground up. Think of this as an opportunity. You can rebuild it an any way you desire.

---

## Consent

*Stub. Needs more content.*

### **How to Have Sex on Purpose [blog post]**

*(the most consent-relevant parts are further into the body)*

A moment to talk about consent. Consent in BDSM is a really big deal, because the stuff we do would be torture without consent. It's sad that it's any different for sex, but not a whole lot of people could convince themselves "well, they seemed like they wanted to be dressed up like a ballerina and smeared with mashed potatoes, they did go up to my bedroom after all" to themselves. You've gotta be sure when you're doing kink. It's not just about having a good experience but about not committing a felony. Wait... isn't that true for sex too? Again. **If you wouldn't punch a person because you were kinda sure they wanted it, don't have sex with them either.** Just be like, "So... wanna fuck?" Gotta tell you, I haven't gotten a lot of "Oh, I was wet and humping your leg and imagining the things I'd do to you, but now that you asked, forget it," from that. I have gotten "no," but thank God for those "no"s! I'm especially glad I asked then!

---

### **#350: Let's crowdsource some feminist sex ed for frat guys. [blog post]**

If someone says no, freezes, pulls back, moves your hands away, goes passive or limp, or seems at all reluctant to do something or less than fully present, doesn't make any moves towards removing clothing, **stop whatever it is you're doing.** Treat "maybe" as "no." Let your partner make the next move, if there is a next move. **Trust that if "maybe" really means "yes," they'll find a way to let you know.**

This might feel awkward and uncomfortable at first because (heterosexual) men are socialized to be the aggressors who must "perform" and move the action along, and women are socialized to be more passive receivers. There's this (bad) cultural expectation that guys

are always up for sex and will be pushy about it and women are gatekeepers and that sex is a favor they do for (or cruelly deny) to men.

Even when people know intellectually that it's bullshit, it's still very possible for that model to feel normal and even good when it plays out in the moment with someone you like. If you deviate from that script, you take a risk that your partner might not step so comfortably into the role of aggressor and that things might unfold more slowly than they otherwise would or require a lot more explicit communication. Trust that the weirdness is momentary. Trust that people who really want you will find a way to make it happen between you – if not Right Now, then soon. And honestly, if your partner is nervous or having second thoughts or worried about being pressured, being No Pressure Guy is the coolest and sexiest thing you can be.

---

## Consent Once Doesn't Mean Consent Always

Things like "consent from one time and one context doesn't imply indefinite further consent" are very important to remember. Things change.

---

## Polyamory

### Polyamory takes what normal relationships do, just more so

Polyamory can be pretty good but I don't think it's always going to be easy. Multiple partners can allow you to get a wider variety of desires met (and have to say no less to things you want) but also jealousy is pretty bad and not trivially solved for everyone.

I think polyamory basically requires all the same skills as needed for monogamous relationships (communication, introspection), but more so – you're playing on hard mode with many people and their expectations/feelings in the mix.

Definitely be reluctant to "open up" long-standing monogamous relationships to poly. I've seen at least one marriage destroyed that way.

---

### More Than Two – Franklin Veaux [Book]

From [Models of human relationships – tools to understand people](#)

This is commonly known as the polyamory bible. It doesn't have to be read as a polyamory book, ~~but in the world of polyamory emotional intelligence and the ability to communicate is the bread and butter of everyday interactions.~~ If you are trying to juggle two or three relationships and you don't know how to talk about hard things then you might as well quit now. ~~If you don't know how to handle difficult feelings or experiences you might as well quit polyamory now.~~

~~Reading about these skills and what you might gain from the insight that polyamorous people have learnt is probably valuable to anyone.~~

Elizabeth: I would strongly recommend removing More Than Two from the post. The primary author has been [accused of abuse](#) by his [co-author](#) and multiple other long term partners (>50% of all long term partners he's had, maybe close to 100%). It's not clear to me his

behavior meets a strict definition of abuse, but you can't get to this stage without some combination of "the relationships were terrible and he contributed to that" and "he has absolutely terrible taste in partners", and I think both are pretty disqualifying for a relationship advice guru. Plus, while I don't have any evidence other than his statements and theirs, the kinds of bad behavior they describe are extremely consistent with the failure modes of what he writes about.

---

## I would recommend [Polysecure](#) over More Than Two [Book]

Attachment theory has entered the mainstream, but most discussions focus on how we can cultivate secure monogamous relationships. What if, like many people, you're striving for secure, happy attachments with more than one partner? Polyamorous psychotherapist **Jessica Fern breaks new ground by extending attachment theory into the realm of consensual nonmonogamy. Using her nested model of attachment and trauma, she expands our understanding of how emotional experiences can influence our relationships.** Then, she sets out six specific strategies to help you move toward secure attachments in your multiple relationships. Polysecure is both a trailblazing theoretical treatise and a practical guide.

---

## Blogposts on Polyamory by Ozymandius

The blog [Thing of Things](#) by Rationalist [Ozymandius](#) has a bunch of relevant posts on polyamory. Probably just search for the best of them. Here are some I could easily pull up (probably not the best ones):

- [On Polyamory Advice](#)
- [You Don't Have To Be Good At Relationships to Be Poly](#)
  - A lot of polyamory advice books are, frankly, terrifying. They make it sound like to be poly you have to be Emotional Competence Georg, who lives in a firm boundary and negotiates with his partners about 10,000 emotional needs each day.

***So I would like to say something reassuring to my crazy friends: you don't have to be good at relationships to be poly. It helps! It definitely helps! The advice in More Than Two or The Ethical Slut is good for people of all relationship styles, monogamous and polyamorous.***

***However, I am needy, whiny, insecure, and approximately as good at communication as a potted plant. And I have been poly for several years and it has worked out fine.*** That's for a bunch of reasons. Polyamory is sometimes easier.

- [Assorted Thoughts on Polyamory](#)
    - An odd thing about polyamory is that you can have your heart broken, be wanting to punch the wall and throw things and curse every time you hear that bastard's name mentioned while simultaneously being bubbly, giggly, happy, full of new relationship energy, tremendously excited by everything about this new person while simultaneously knowing that your rock is there, your [secure base](#), who will always be there for you if you need them.
-

# Providing Support & Expressing Love/Affection

## Love Languages (aka how to express and receive affection effectively)

**Love languages is a neat concept - the way people experience and express affection can be different**, so it's good to have a good model of your partner and what really reaches their heart.

Classically there are five love languages: words of affirmation, acts of service, touch, quality time, and gifts. In fact, I think there are many more.

One thing some people really care about is being "seen", having someone understand their experience and anticipate their needs and desire. For others, it's feeling "wanted".

Personally, I realized relatively recently that playful teasing (or outright [outrageous countersignalling](#)) is important to me for feeling safe and comfortable and connected to someone; it really is one of my love languages.

---

## Related to the idea of Love Languages is [Personal User Manuals](#).

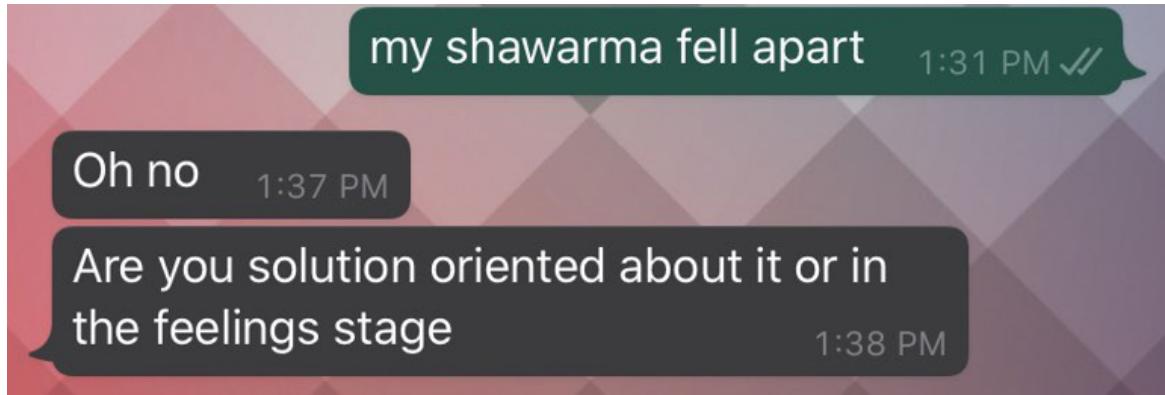
It's a concept I learned in the workplace, but it should generalize. For your friends, partners, etc., write up a document that explains your default personal culture and preferences: how you like to communicate, what makes you happy/unhappy, etc. etc. Seems worth doing for relationships.

---

## Learn to Listen: Problem-Solving vs Support

This is classic advice but just always worth remembering. At different times and across different people, partners want different things from conversations. Model them!

A couple I know actually defined between themselves a few modes of nuanced support so they could say things like "Would you like support-style A or B right now?"



## Warnings & Cautions

### **Being a savior is risky / Trying to fix others is risky**

It's a not uncommon pattern for someone to see someone they're interested in struggling with a particular problem and think they can help them solve it. This is risky. Mixing your interest in someone with a desire to help them...it's tempting but I think sets up bad dynamics. It might involve escalating them while they're in a vulnerable state, it might cause them to end up feeling obligated to reciprocate romantic attention when they don't want to, or very likely, you're not actually in a good position to help them and understand what's going on less well than you think.

Six or seven years ago, early in one relationship the person I was with seemed to be struggling with psychological challenges I myself didn't have, so I thought I could just easily impart how I approached those topics and thereby fix her. But I didn't really understand and so instead I made it so she didn't want to talk to me about her challenges for a really long time. I just didn't actually understand.

In another relationship, the person had not that long ago left a very abusive relationship. I thought that I could be the complete opposite - loving, caring, considerate. Except that I didn't actually understand how she felt or what she needed at that point, so my well-intentioned caring actually missed the mark and made her feel worse in many ways.

This isn't to say you shouldn't try to help others, but be careful when you're combining it with your romantic interest.

The other point to remember is that you can't really fix other people, definitely not despite themselves. You can at best help them help themselves, and if they don't want that, there's not obviously much you can do.

### **Probably don't make your relationship contingent on the other person changing**

Sometimes you'll meet someone who you think you could potentially like if they were different in this one important way, if they just improved a little (or a lot), and you think you

can help them make those improvements. I won't say this is never true, but it's an anti-pattern, for sure.

There's a kind of crazy book, [The Mastery of Love: A Practical Guide to the Art of Relationship](#), that nonetheless has some spirit of wisdom to it:

"You cannot change other people [not literally true, but ok]. You love them the way they are or you don't. You accept the way they are or you don't. Try to change them to fit what you want them to be is like trying to change a dog for a cat, or a cat for a horse. That is a fact. They are what they; you are what you are. You dance or you don't dance. You need to be completely honest with yourself - to say what you want, and see if you are willing to dance or not. You must understand this point, because it is very important. When you truly understand, you are likely to see what is true about others, and not just what you want to see."

The passage is literally true and loving someone because you want to grow alongside them and become more is great, but also there's some spirit in that passage that feels right to me though it didn't at first.

---

## Have other relationships! Diversify!

Standard failure mode, especially for busy people, is to invest in their romantic relationship and neglect other social connections. Naturally a bad idea because you've got a single point of failure. What do you do if you're having a rough patch with your partner? What do you do if they're away or busy? What do you do if it ends?

Even when you're busy and super excited about your partner, nurture other friendships.

---

## Thinking you're good at relationships is a recipe for not noticing when you're being bad at relationships

Believing that you are X, e.g. kind and thoughtful in relationships, and particularly priding yourself on it, is a surefire way to fail to notice some occasions when you are not being X.

Examples: empathetic, considerate, "rational", "I confront my problems head-on rather than avoid them".

Such failures to notice will be consequential.

Well, I dunno, has been true of me. I assume I'm not *that* special and this applies enough to enough others to be worth the warning. For a long time I've tried to pride myself less on various traits and behaviors, but therein lies the meta-failure I very much have committed: by believing I have reduced my pride, I fail to notice where I haven't.

Due to the above mechanics, I expect to be caught out failing to instantiate various traits that I valorize and have laid claim to. Not because I don't really want or value having the traits, just because noticing failures is hard. You can call me out on things when I'm not being who I like to believe or say I am; I will endeavor to be glad rather than mad. I hope you will forgive me if in my failures I have wronged you. - [FB post](#)

---

# Sex

*How to have good and healthy sex is beyond the intended scope for this thread, but I welcome people to add links to external resources here (or submit them via comments with spoiler text/warning, or the [Google Form](#)).*

Let's just assume these are all NSFW (links hidden behind spoiler text cover)

- [The pragmatist guide to sexuality](#) (lots of data on what people's sexuality is actually like! 100% recommended (though maybe just skip the authors interpretations and look at the tables))
  - [The Typical Sex Life Fallacy](#)
  - [What's it like to have sex with Duncan?](#)
    - actually will help you reflect on health/unhealthy/good/bad sexual interactions
  - Aella (on Twitter and elsewhere) has lots of interesting surveys about sex stuff, particularly kink/fetishes
- 

## Models and Concepts

### Attachment Styles

Attachment styles is this relationship model that generalizes from infant attachment studies to adult relationships styles. The inference feels like wonky science, but I still feel like there's something to it. Quick summary of attachment styles:

Secure: a securely attached person feels comfortable and at ease-with the relationship. They seek and respond to bids for attention/affection in a way that's healthy for both partners.

Anxious-preoccupied: is anxious about the relationship and is preoccupied seeking reassurance about the relationship.

Fearful-avoidant: although they want the relationship on some level, they are afraid of the intimacy and "pull away"

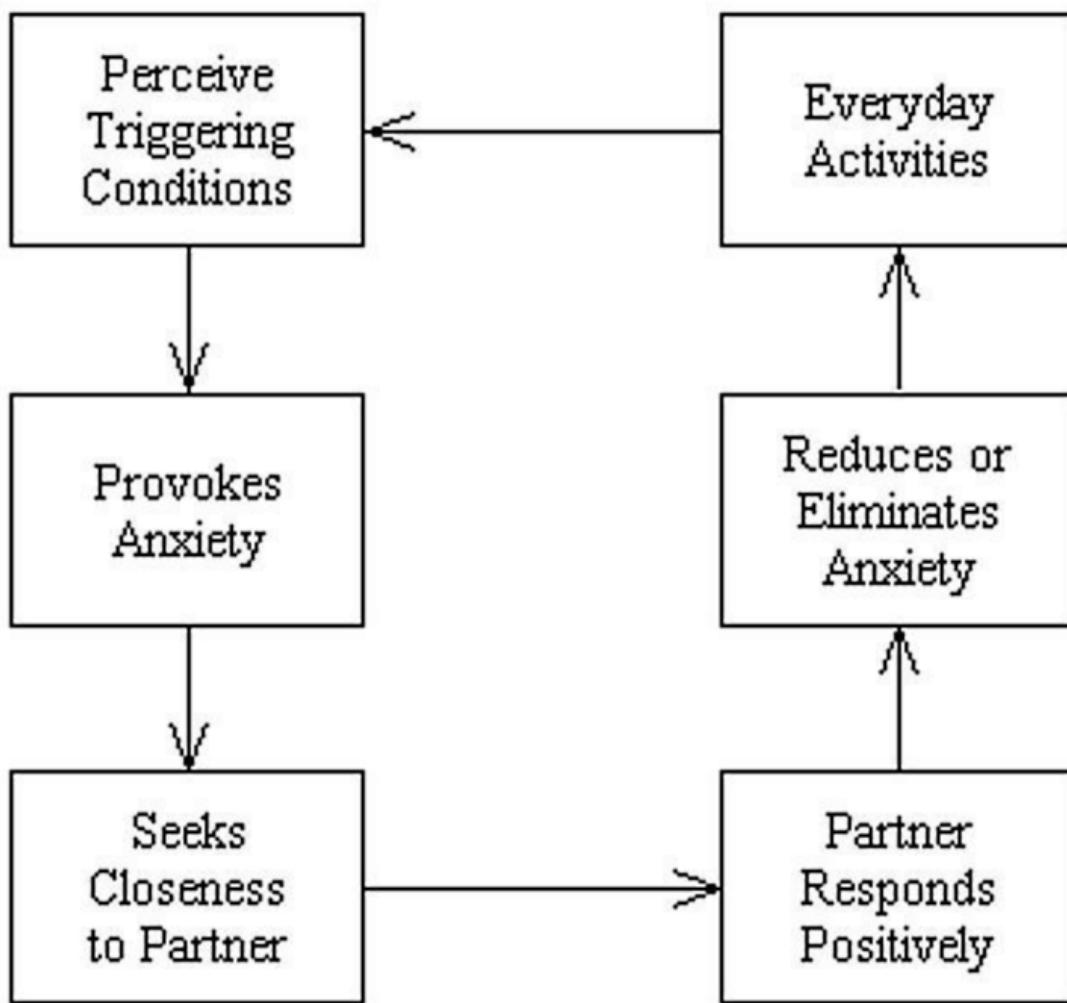
Dismissive-avoidant: like the fearful-avoidant but more extreme, while they vaguely want the relationship they also kind of believe they don't need it and put up a lot of walls.

Anxious-preoccupied and the avoidants classically make for a bad dynamic. In one order or the other, you have the anxious pre-occupied pursuing reassurance and the avoidants pulling away, thereby inducing more anxiety in the anxious-preoccupied who pursues reassurance further, inducing more avoidance, etc.

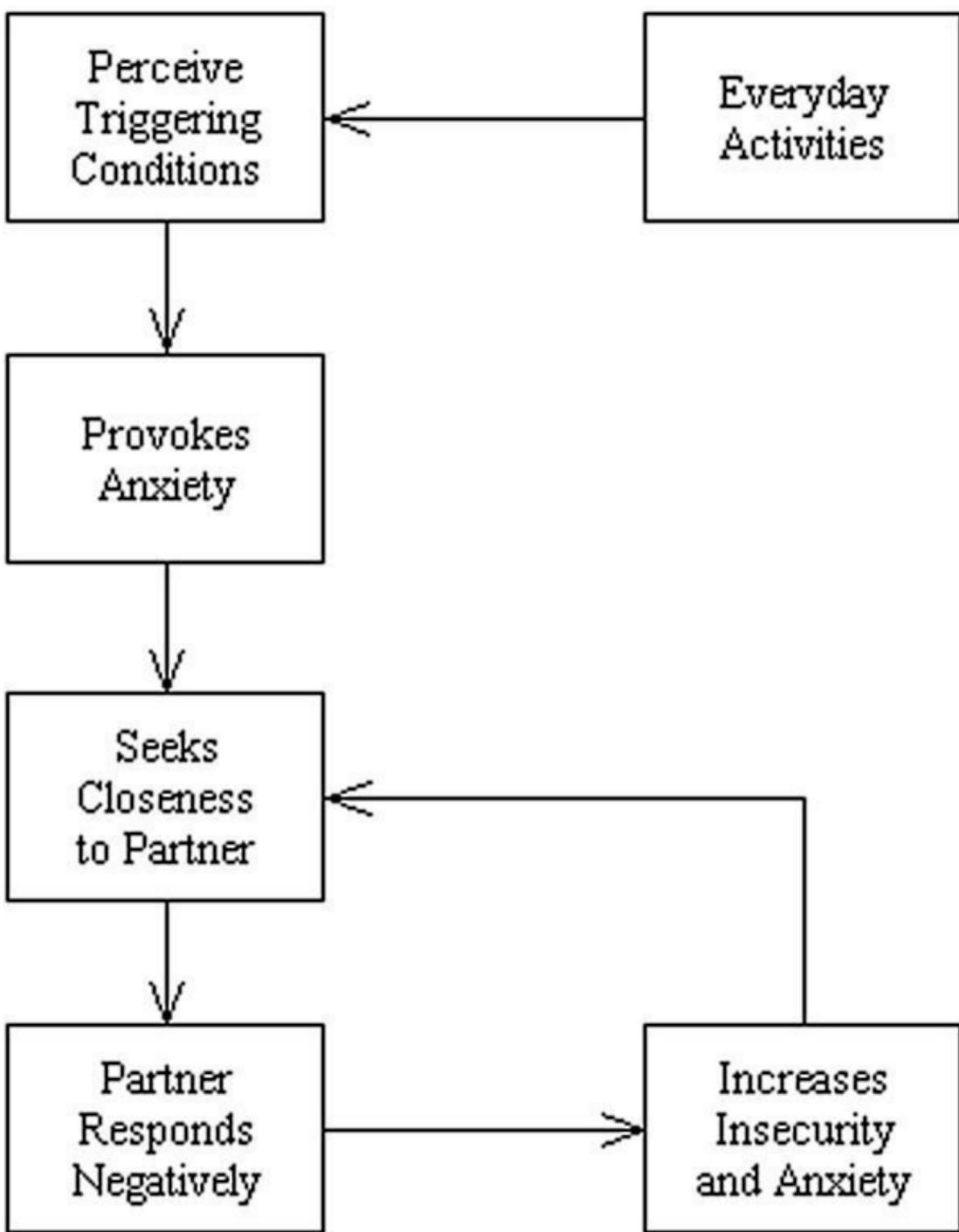
These are more "patterns" of behavior than essential traits of people. Anxious behavior in one person can induce avoidance behavior in an otherwise secure person, and vice versa.

There are several pitfalls here, but one is that you tell an anxious-preoccupied person about attachment styles so they say "ok, I will inhibit my anxious behaviors so they will like me". This will likely fail because it comes from the same frame of desperately needing to be liked.

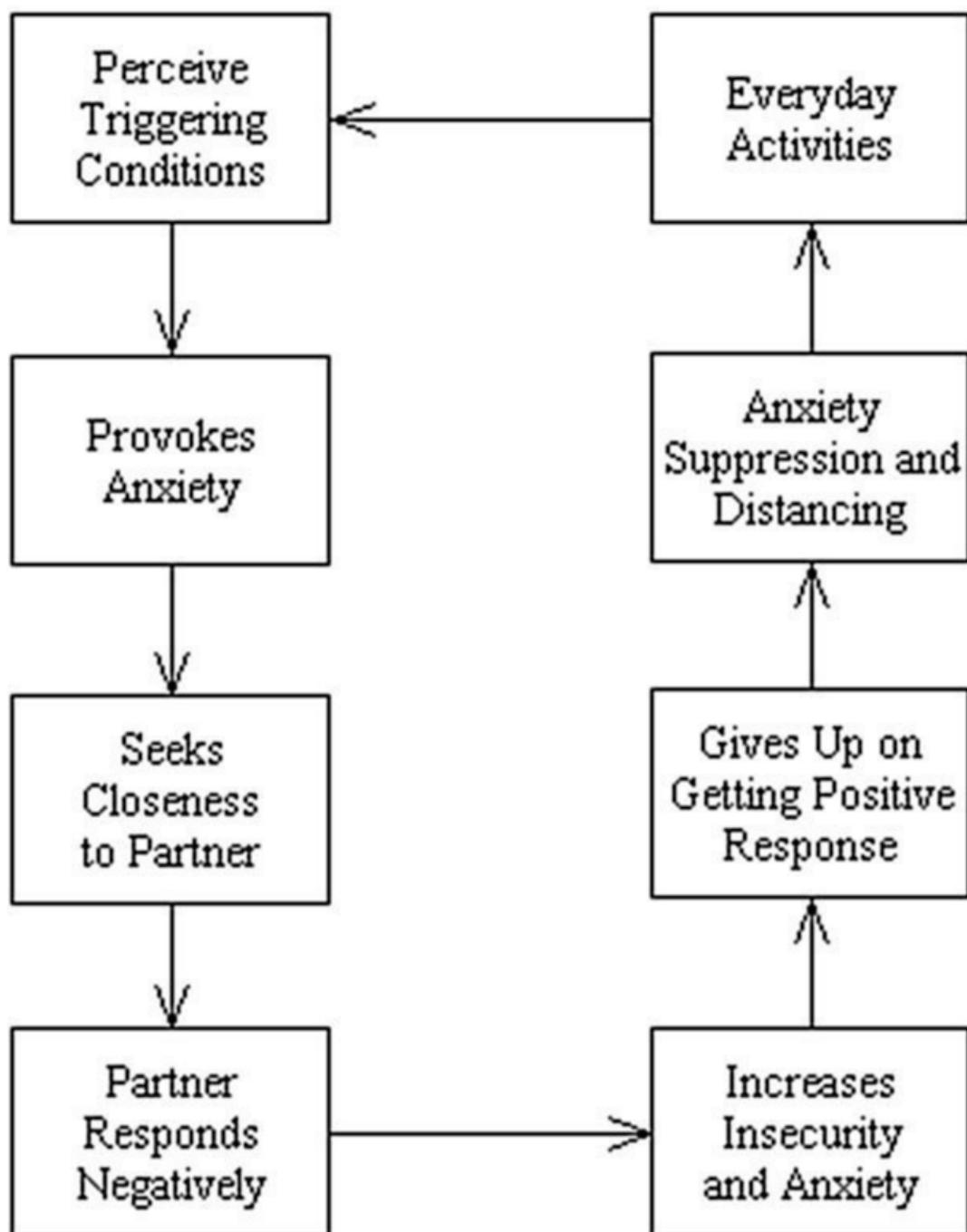
The flowchart for the **security-based** strategy:



Secure Attachment Pattern



Anxious pattern



Attachment-Avoidance pattern

[Images from How to Love \(or Leave\) a Dismissive Partner](#)

## Power Differences

## Power dynamics between people in EA [excerpts]

### **It's harder for junior people to give their real opinion**

Opinions that junior people express will often be shaped (maybe unintentionally and unknowingly) by what they perceive senior people's opinions to be or what they think the senior people want to hear.

Possible steps:

- Ask junior people to share thoughts before you give yours. Listen openly and show interest.
- Give them more time and encouragement to lay out messy thoughts.
- Set the stage for welcoming messy / unformed thoughts by sharing some of your own.
- Give encouragement and appreciation when they offer opinions and especially when they disagree with you - make it a good experience for them.
- Point to any tangible changes made based on critical feedback, so people can see you take it seriously.

### **It's harder for junior people to establish boundaries**

A senior EA I know notes that she's effusive with hugs at work. She realized after a long time that one of the junior staff she was working with prefers less physical contact than most people, and that she was probably making this person uncomfortable by offering hugs that they didn't feel comfortable turning down.

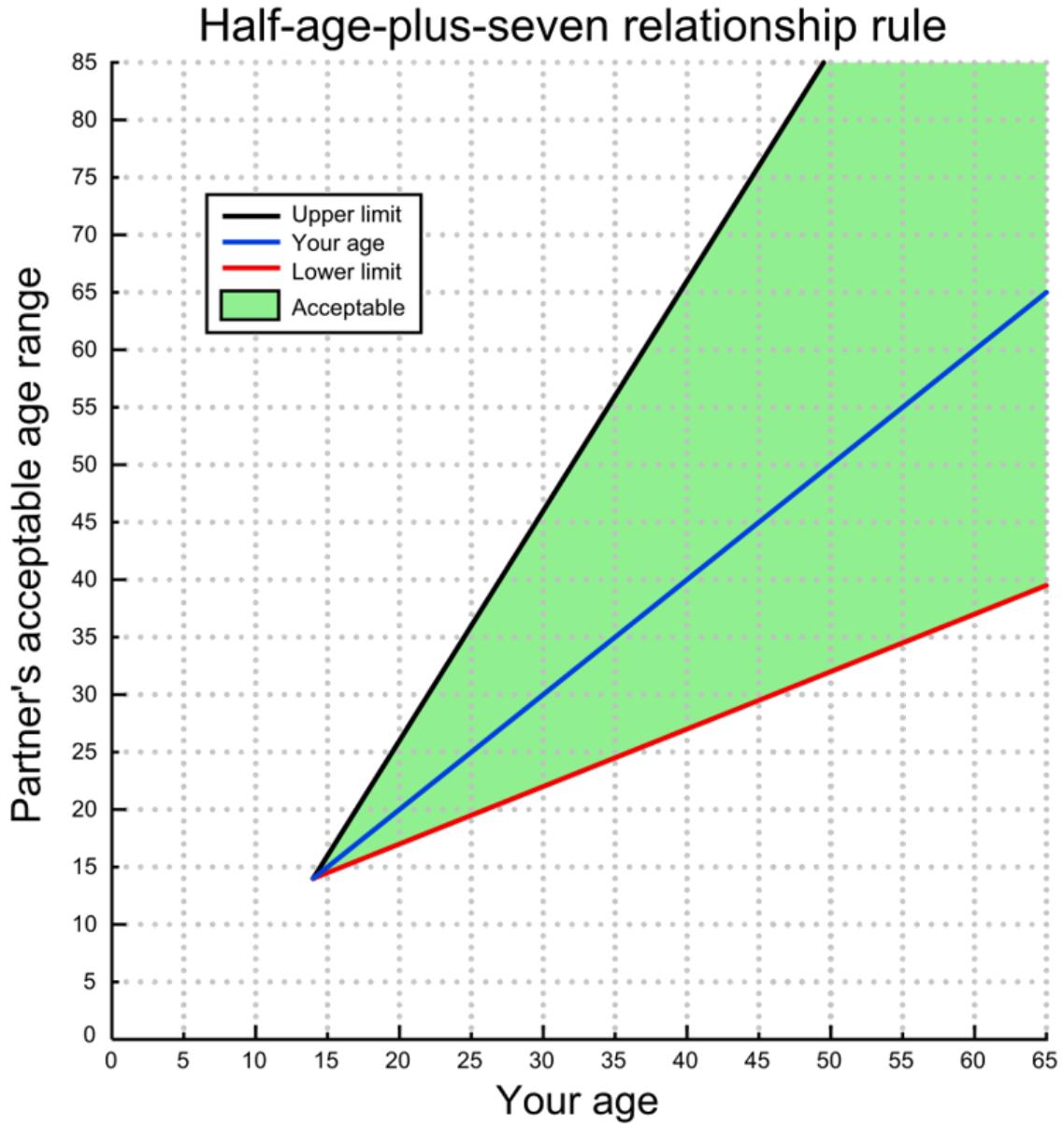
### **Social situations are also assessments**

When junior people are in the same space with senior people, they often correctly feel that they're being evaluated as potential future grantees or hires. Every lunch feels a bit like a job interview. The good aspect of this is that they would probably like to have a foot in the door, but the bad aspect is that it can make time in EA spaces pretty anxious because even minor social interactions feel high-stakes.

---

## **Age Differences**

*Half your age plus seven - common wisdom*




---

## Thoughts on dating younger people

Regarding dating younger people, I think there are some things which bump up the usual risk of relationships being rough, maybe by like 2x or 3x (so if you have a 3% chance of it being Quite Bad and a 12% chance of it being Normal Bad, you pop up to maybe 9% and 36%, respectively).

Like, I think that the only things that make it bad are the things that make *any* relationship bad: abuse of power differentials, coercion or mind-bending/gaslighting, selfishness of various kinds, failure to communicate, etc.

I think the most important things to track:

1. Try to note if there are Big Traumas or triggers in either of you; the main thing that sent me and one partner in the wrong direction was me not finding out in the early and mid days the sheer extent of her abusive upbringing.

2. Know the difference between your own preferences<sup>[1]</sup> and your own preferences, like which things you Really Want versus which things you Kinda Want But Only If She Wants Them, and vice versa. Like, things get askew when asks are not really "just asks" or whatever; having clear selfsight there goes a long way.

Standard don't-be-level-one-stupid advice on age gaps is, imo, level two stupid. The standard advice is, like, the older or more powerful person should be the "bigger" person, and be more willing to sacrifice and so forth. To take their own needs as object, create more space for the younger person. And this is absolutely true, but as a *side effect*. If you target that, you fall prey to a weird Goodharting thing where you're always the "grownup" in the room and you're not allowed to want things for fear of creating pressure, etc.

I think the don't-be-level-two-stupid thing is know your own wants, needs, and desires. Know their relative strengths, and how eagerly/pressured-ly you are motivated by them. Be clear about what you want (not necessarily all right up front/first date, but reasonably early on) and seek the overlap. Be able to see where you can get A, B, and D from this partner, but not C or E, and then be smart about that. Like either get C and E elsewhere, or genuinely make peace about it, or whatever. Avoid fabricated options.

Put another way: I think people should be, and are, motivated *both* by wanting genuinely good things for their partner, *and* by wanting genuinely good things for themselves. Where people run into trouble is where they can't tell the difference between these two buckets. They talk themselves into "this thing I really want just for my own sake is *good* for them" instead of, y'know, asking or checking, and believing the info that comes back. With age gaps or power differentials, it's easy to accidentally overwhelm the less-experienced partner, if you're not watching yourself for that. (So I claim.)

---

## Thoughts on age gap relationships

Things that give me a baseline skepticism of age gap relationships including some examples of things going wrong:

- very different social / financial stability-power, which leads to the younger person making more compromises or feeling more insecure
- being at different life stages, and the younger person missing important opportunities or feeling pressured to do things they're not quite ready for due to trying to sync with the older person's needs (but I think this doesn't apply as much for non-primary relationships)
- **younger person has less relationship experience and doesn't realize what things are negotiable/normal, leading to them advocating less for their needs/wants than you'd expect**
  - "what things are negotiable / normal" = concrete example from my life, not of bad behaviour, but it was like 4ish years into my sex life before a partner was like "hey, uh, you can say 'no' to sex whenever and I'll still like you" and I apparently really needed to hear that spelled out explicitly?? similarly, I did some dubious-consent stuff with a partner before realizing that just because someone is a dude does not mean they are always DTF in whatever configuration and you should also check in... I feel like it would just be super easy to miss stuff like this if you've been dating people with more relationship experience for a while? but also this shows up in, like, "when is it okay to talk about jealousy" or "how often are you supposed to see a partner". I was talking with a partner about this (in a general pattern way) and he used the metaphor of how travelling to a new

country is supposed to open your eyes to all these ways you assumed things were simply done, but actually that was just your culture, what! I feel like the same sort of learning can apply to having multiple relationships?

- (this is my age gap failure mode that sucked) **older person assumes that the younger person is implementing a certain relationship interface** (e.g. "we had a talk where you said you were okay with us not being committed, so I'm assuming this is fine"), doesn't check in enough that the underlying assumptions are true and is careless/hurtful as a result (there's a difficult balance here where you want to respect the younger person's agency but also need to be extra careful)
  - I had one relationship where I was like "cool, we had a really explicit expectations-setting conversation, she said she was fine with this arrangement, so she probably is" when in fact the situation was more like "she is accepting a on-net-bad-for-her relationship because she doesn't feel like she can negotiate for anything else" and I think that if I had been more carefully modelling her as less relationship-experienced I would not have been so casual about "cool we set up this relationship check-in where she could bring stuff up and she didn't, therefore all is chill"
- **younger person overweights older person's opinions / mixing mentorship and romance, leading to them signing onto things they otherwise wouldn't**
  - I know someone who dropped out of school to do a coding bootcamp on an older partner's recommendation, and this was not a good idea in the end, and maybe they would have given it more independent thought / trusted the suggestion less if their partner hadn't been older? hard to know, tho
- **precocious young person really wants to be taken seriously as an adult, overrides their own boundaries in order to appear more mature** by what they perceive their partner's definition of that to be (I've seen this lead to two cases of quite severe abuse)

some things I think you can do here:

- precommit to helping them break up with you if need be (my first girlfriend, 18 to my 16, talked me through breaking up with her; one of my partners talked a younger partner through breaking up with him)
- since they don't know about the heterogeneity of relationship culture, spend more time asking questions along the lines of "hey, just so you know, it's okay if you [---]?" or "just to check, would you rather [----]?" I want to make sure we don't just fall into [.....] because that's how I often do things"

---

## Inexperienced people can pick up bad habits from dating other inexperienced people

There's another risk that a person who's been dating romantically **inexperienced/unskilled people (e.g. likely other young people) might have picked up some bad habits**. For example, a good habit is that when you're unhappy/want to say no/change your mind/etc, you say this explicitly and proactively; however, if your past partners have responded poorly when you've done that, you'll acquire the habit of not actively expressing anything that might get a negative reaction.

---

## If you're unhappy in your relationship

If your relationship makes you unhappy, do something. The unhappiness is a sign that something should change.

I was once in a relationship where I was extremely attached to the person, though I was miserable due to some bad dynamics we had going on. The thought of ending it was unbearable, but I was also really unhappy. I kept thinking to myself "If I'm so rational, why do I feel so trapped in this bad situation?" Eventually, I got myself out of it, but I'd like to think that I won't get myself into it ever again. I hope I've learned to really sit up and act when I'm that unhappy.

## Abusive Relationships

*Please help fill this section out!*

### It's not normal to be harmed

**A classic and maybe even defining feature of abuse is that the abused person is made to feel that it is normal or even right for them to be harmed.** They're told "You deserve it." Or "this is just what relationships or families are like." Or "you aren't being harmed, you're fine." Over time, abused people may come to believe this.

---

### Don't let the fear of leaving trap you

I don't personally have direct experience of abusive relationships but I think I've observed a few things at a distance: **people end up trapped in abusive relationships because even though they're unhappy, the thought of leaving feels worse.** And that's how you end up trapped. I think the escape is realizing that while leaving will be painful, it's survivable and is for the best.

The other thing I've observed is people staying in relationships they don't like out of fear of angering their partner. I don't know how to solve that, but I just want to say "**don't let their anger control you!**" You deserve better.

It's likely a very large red flag if either you feel embarrassed to describe your relationship to others or if your partner would get mad at you doing so.

---

### Privacy norms can be wielded as an obfuscating weapon

Sometimes, privacy is wielded as a tool to enable manipulation.

I've run into a couple people who exploited my good faith / willingness to keep things confidential, as part of an overall manipulative pattern. Unfortunately, I don't feel comfortable going too far into the details here (please don't speculate in the comments), which makes it a bit harder to sanity check.

---

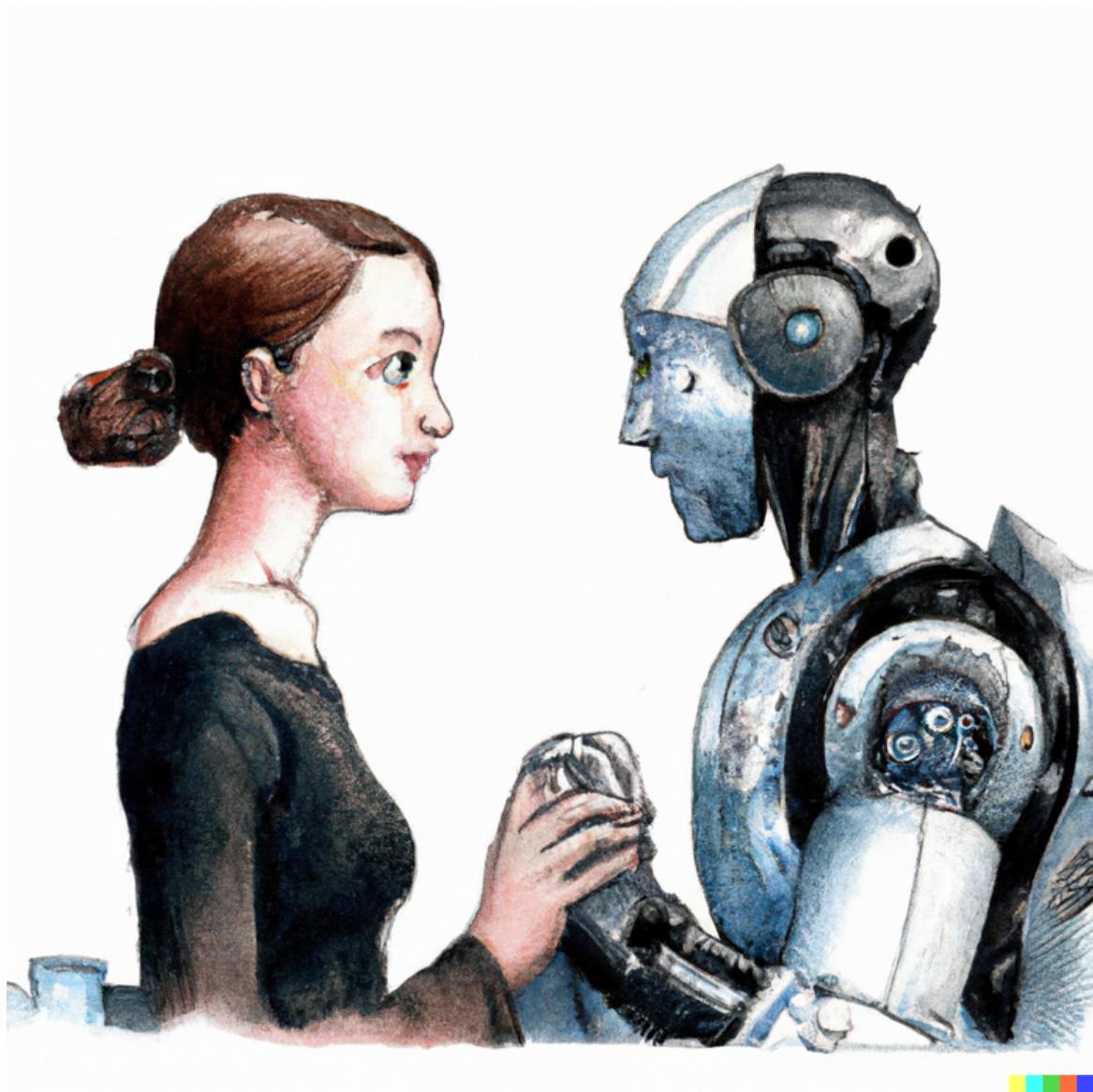
## Resources

*Probably organize this section better if it grows to have more stuff. Anyone feel free to sort as they feel fit.*

- [LessWrong Relationships tag](#)
- [LessWrong Communication Cultures tag](#)
- [Reciprocity.io](#)
  - Rationalist/EA dating site where you mark interest in people and two people are only notified if they both expressed interest
- [Avoidant: How to Love \(or Leave\) A Dismissive Avoidant Partner \[book\]](#)
  - Generally helpful book on the topic of Attachment Styles. Useful for understanding that different people have learnt (or deeply ingrained) different patterns of behavior in relationship and that the interaction of these patterns matters.
- [The Mastery of Love: A Practical Guide to the Art of Relationship \[book\]](#)
  - This is a crazy book. But has some great stuff on letting your partner be who they are recognizing that you're not responsible for solving all your partner's problems.
- [How to Talk So Kids Will Listen and Listen So Kids Will Talk \[book\]](#)
  - A surprisingly helpful book for adult-to-adult communication
  - [Mo Nastri](#): I also enjoyed weft's [Book Review: How To Talk So Little Kids Will Listen](#), written by Julie King and Joanne Faber, daughter of Adele Faber, who co-wrote the former with Elaine Mazlich. Quoting weft:
    - The core principles are the same, but the update stands on its own. Where the original "Kids" acts more like a workbook, asking the reader to self-generate responses, "Little Kids" feels more like it's trying to download a response system into your head via modeling and story-telling. I personally prefer this system better, because the workbook approach feels like it's only getting to my System 2 (sorry for the colloquialism). Meanwhile being surrounded with examples and stories works better for me to fully integrate a new mode of interaction.
    - I too prefer examples and stories to self-generated responses, so I thought it'd be a useful complement to others like weft and I.
- [Elizabeth](#): I found [Crucial Conversations](#) to be the adult version of How To Talk So... and it seriously levelled up my interpersonal skills at the time.
- [Mark Manson Relationship Advice](#)
  - Came up on a first Google search, pretty reasonable stuff
- [20 People on the Best Relationship Advice They Ever Received](#)
  - Another Google result that seems pretty reasonable
- [That's Not What I Meant! How Conversational Style Makes or Breaks Relationships \[book\]](#)
  - Haven't read it but maybe good?
- Why We Love: The Nature and Chemistry of Love [book]
  - Read it five years ago and don't really remember it. Focuses on chemical/evolution aspects and makes interesting comparisons of love to addiction. Maybe of interest to the especially limerent
- [Captain Awkward Dating Guide for Geeks](#)
- [The Ferrett](#) - blog with a lot of great, thoughtful posts about poly and relationships
- [Reddit: Am I the asshole?](#)
  - People post about conflicts and get the crowd's opinion of who is being unreasonable
- [Models of human relationships - tools to understand people \[review post\]](#)
  - Summary/review of a long list of relationship books
- [Morpheus](#): I'd add [the pragmatist guide to relationships](#) (has material on seeking as well as on maintaining relationships) I read like half the book (the parts remotely relevant to me (the book has ~660 pages)) and is very much written from a more selfish livehacking/"munchkinism"/economist (markets and contracts) kind of perspective, which I found entertaining, but which might be off-putting for some. The authors also know and seem to practice their Bayesian epistemology, and the book

held up pretty well to online spot checks, and asking people. I still felt like sometimes the authors didn't add enough uncertainty disclaimers around their theories about humans, but it's not like I wouldn't have similar complaints about some Iw posts. and

- [Plays Well with Others: The Surprising Science Behind Why Everything You Know About Relationships Is \(Mostly\) Wrong](#)
- [Books I read 2017 - Part 1. Relationships, Learning \[review post\]](#)
  - same author as last one, mostly the same books too
  - *post contains brief reviews/descriptions of each book*
  - **Relationships & Communication**
    1. [Having Difficult Conversations - Douglass Stone](#)
    2. [Crucial Confrontations -Kerry Patterson](#)
    3. [Emotional Intelligence - Daniel Goleman](#)
    4. reread: [How to Win Friends and Influence People](#), circa 2007 - Dale Carnegie
    5. [More Than Two - Franklin Veaux](#)
    6. [Nonviolent Communication - Marshall Rosenberg](#)
    7. [Living Non-Violent Communication - Marshall Rosenberg](#).
    8. [Daring Greatly - Brene brown](#)
    9. [On Apology - Aaron Lazare](#)
    10. [Circling Handbook - Marc Beneteau](#)
    11. [7 Principles for Making Marriage Work - John Gottman](#)
    12. [Feeling Good Together - David D Burns](#)
    13. [Getting Past The Pain Between Us - Marshall Rosenberg](#)
    14. [Graduating From Guilt - Holly Michelle Eckert](#)
    15. [The Surprising Purpose of Anger - Marshall Rosenberg](#)
    16. [Come as You Are - Emily Nagoski](#)
    17. [Jono Bacon - The Art of the Community](#)
    18. [Games People Play](#)
    19. [The Stories we tell Ourselves](#)
    20. [Sex at Dawn](#)
      1. Elizabeth: Sex at Dawn is also atrocious from a scientific perspective, although much less likely to cause overt harm [than More than Two].



## 1. ^

Coference is a preference that is referenced on someone else's preference, e.g. "I have a preference for your preference or something that's the combination of our first-order individual preferences.

# Yes, AI research will be substantially curtailed if a lab causes a major disaster

There's a narrative that [Chapman](#) and other smart people seem to endorse that goes:

People say a public AI disaster would rally public opinion against AI research and create calls for more serious AI safety. But the COVID pandemic killed several million people and wasted upwards of a year of global GDP. Pandemics are, as a consequence, now officially recognized as a non-threat that should be rigorously ignored. So we should expect the same outcome from AI disasters.

I'm pretty sure I have basically the same opinion and mental models of U.S. government, media, and politics as Eliezer & David, but even then, this argument seems like it's trying too hard to be edgy.

Here's another obvious historical example that I find much more relevant. U.S. anti-nuclear activists said for years that nuclear power wasn't safe, and nuclear scientists replied over and over that the activists were just non-experts misinformed by TV, and that a meltdown was impossible. Then the Three Mile Island meltdown happened. The consequence of that accident, which didn't even conclusively kill any particular person, was that anti-nuclear activists got nuclear power regulated in the U.S. to the point where making new plants is completely cost inefficient, as a rule, *even in the event of technology advancements*.

The difference, of course, between pandemics and nuclear safety breaches, is that pandemics are a *natural phenomenon*. When people die from diseases, there are only boring institutional failures. In the event of a nuclear explosion, the public, the government, and the media get scapegoats and an industry to blame for the accident. To imply that established punching bags like Google and Facebook would just walk away from causing an international crisis on the scale of the Covid epidemic, strikes me as confusingly naive cynicism from some otherwise very lucid people.

If the media had been able to squarely and emotively pin millions of deaths on some Big Tech AI lab, we would have faced a *near shutdown* of AI research and maybe much of venture capital. Regardless of how performative our government's efforts in responding to the problem were, they would at least succeed at introducing extraordinarily imposing costs and regulations on any new organization that looked to a bureaucratic body like it wanted to make anything similar. The reason such measures were not enforced on U.S. gain-of-function labs following Covid, is because Covid did not come from U.S. gain-of-function labs, and the public is not smart/aware enough to know that they should update towards those being bad.

To be sure, politicians would do a lot of other counterproductive things too. We might still fail. But the long term response to an unprecedented AI catastrophe would be a lot more like the national security establishment response to 9/11 than it would our bungling response to the Coronavirus. There'd be a TSA and a war in the wrong country, but there'd also be a DHS, and a vastly expanded NSA/CIA budget and "prerogative".

None of this is to say that such an accident is likely to happen. I highly doubt any misaligned AI influential enough to cause a disaster on this scale would not also be in a position to just end us. But I do at least empathize with the people who hope that whatever DeepMind's cooking, it'll end up in some bungled state where it only kills 10 million people instead of all of us and we can maybe get a second chance.

# Units of Exchange

**Epistemic status:** Established and confirmed

*The lessons taught in Units of Exchange are straightforward applications of extremely well-established principles from economics and sociology, such as supply and demand, Pareto curves, value of information, the sunk cost fallacy, and arbitrage. The causal relationships underlying by each of these principles have been robustly confirmed in a wide variety of domains, and the recommended actions we've derived from them are simple and fairly conventional.*

---



*"Aw, \$20? I wanted a peanut."*

*Twenty dollars can buy many peanuts.*

*"Explain how."*

*Money can be exchanged for goods and services.*

A version of this course is taught near the beginning of every CFAR workshop, often as the very first class. That's not because the concepts it covers are revelational or groundbreaking, but rather the opposite—they're core concepts, fundamental prerequisites that underlie and inform much of the rest of our content.

If you're already familiar with them—great! This is a quick-and-dirty overview—we don't mean to condescend to people who've already had specific training in these fields, only to provide those same tools to all of our participants. If these are not the droids you're looking for, feel free to skip ahead to Inner Simulator.

---

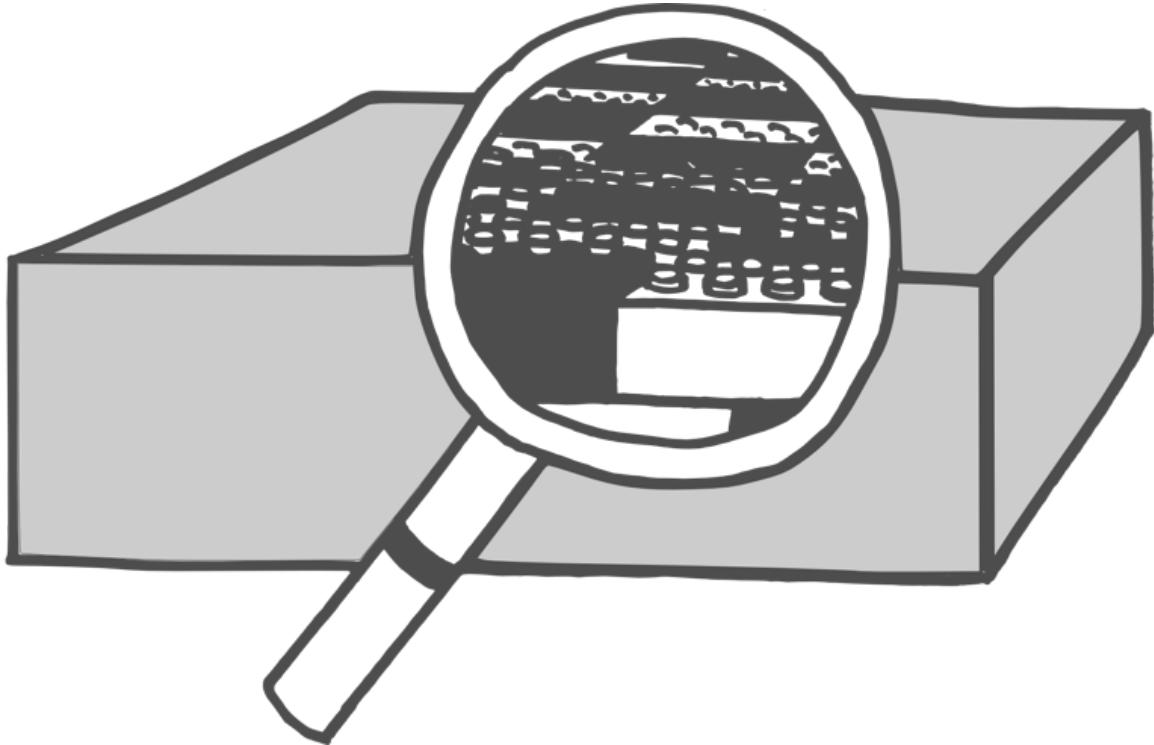
## The Lego Principle: Bricks and bargains

If you've ever done proofs in math or logic, then you know that it's possible to reach complex and interesting results by starting from a limited number of assumptions. The conclusions in Units of Exchange aren't mathematically rigorous, but they do emerge from two key premises, which combine to form what we call the Lego Principle.

### 1. Things are made of parts

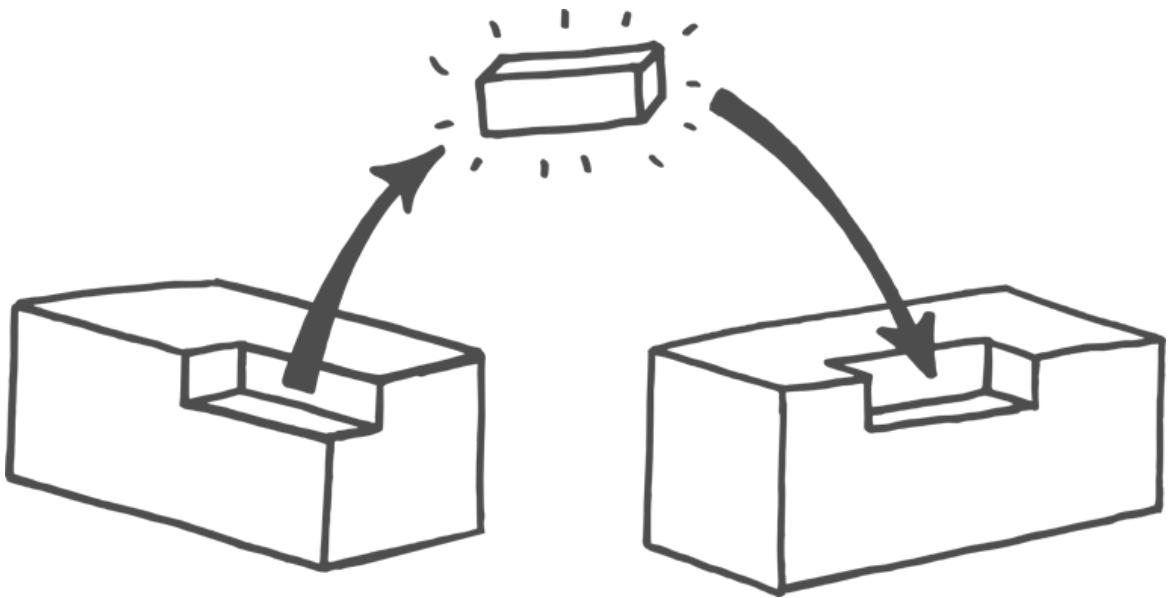
The first half of the Lego Principle is **reductionism**, or the idea that, having fully explained all of the components of a thing and how those components interrelate, there's nothing left to say. Metaphorically, if one has described the trees, shrubs, and fauna in all of their relevant detail, one has *fully explained* the forest; there is no ephemeral "missing" property that is forest-ness.

Reductionism is a powerful concept, because it allows us to interface with complex phenomena by dealing with smaller, simpler sub-phenomena. Brains, emotions, societies, financial markets—these are all large, and sometimes daunting to engage with. But neurons, cognitive if-then patterns, social norms, and individual commodities are all at least *relatively* more tractable.



## 2. Parts may be exchanged

The technical term here is **currencies**, and the idea is that, just as we can trade dollars for pounds (and buy things with either, in a place like an international airport), so too can we trade money for effort or sleep for knowledge or respect for social influence. In particular, where things are made up of *similar parts*, we can make exchanges between them, swapping our resources around to prioritize what we think is important.



Most of the rest of this unit boils down to straightforward applications of these two premises. There are things that we want more of—time, money, motivation, pleasure, attention, energy, knowledge, stuff, sleep, respect, belonging, accomplishment, the well-being of the people we care about. Some of these wants are instrumental—we want them because they will lead to other good things (money being the classic example). Others are more terminal—they’re good in and of themselves (such as happiness or satisfaction).

Since life isn’t perfect and most of us aren’t all-powerful, we make trade-offs. We skimp on sleep to get more work done, bail on a work party to spend time with a significant other, skip dessert because we’re trying to get in shape. To a great extent, making good decisions can be framed as paying close attention to the exchange rates between these various currencies.

\$\$

Money



Time



Energy/Willpower



Affection/Goodwill



Attention



Knowledge



Pleasure

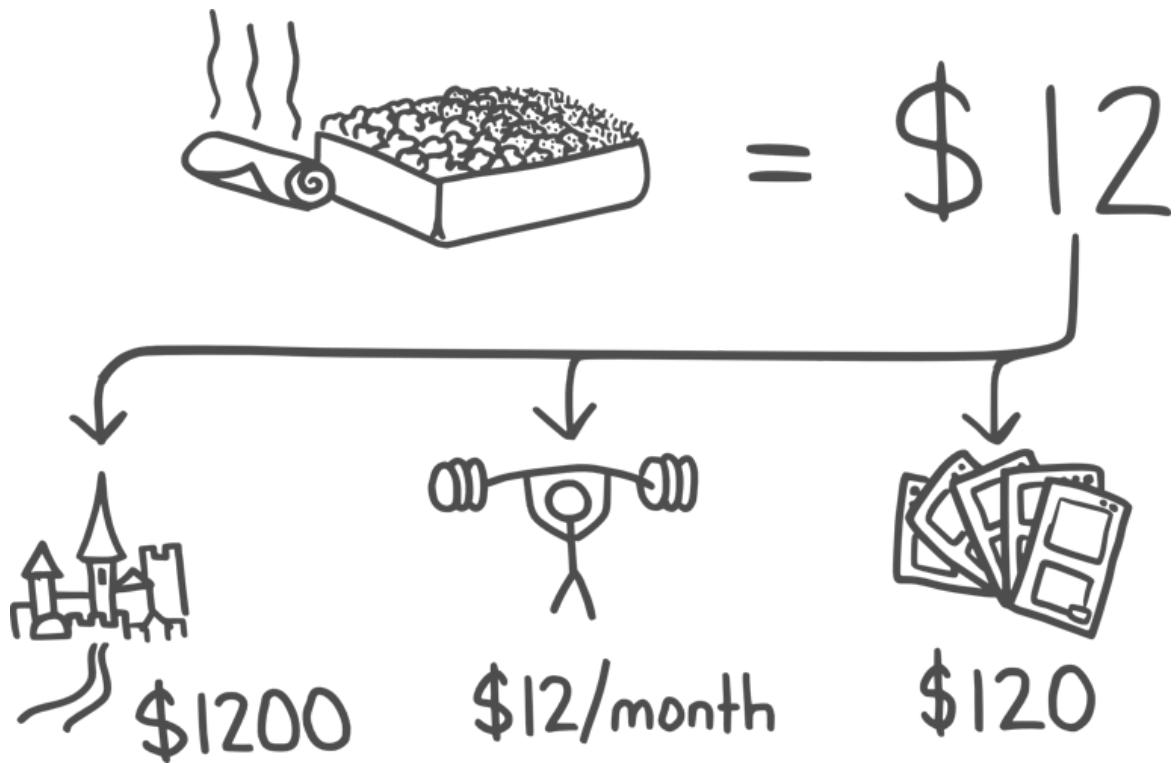


Rest

---

## Part I: Apples to oranges

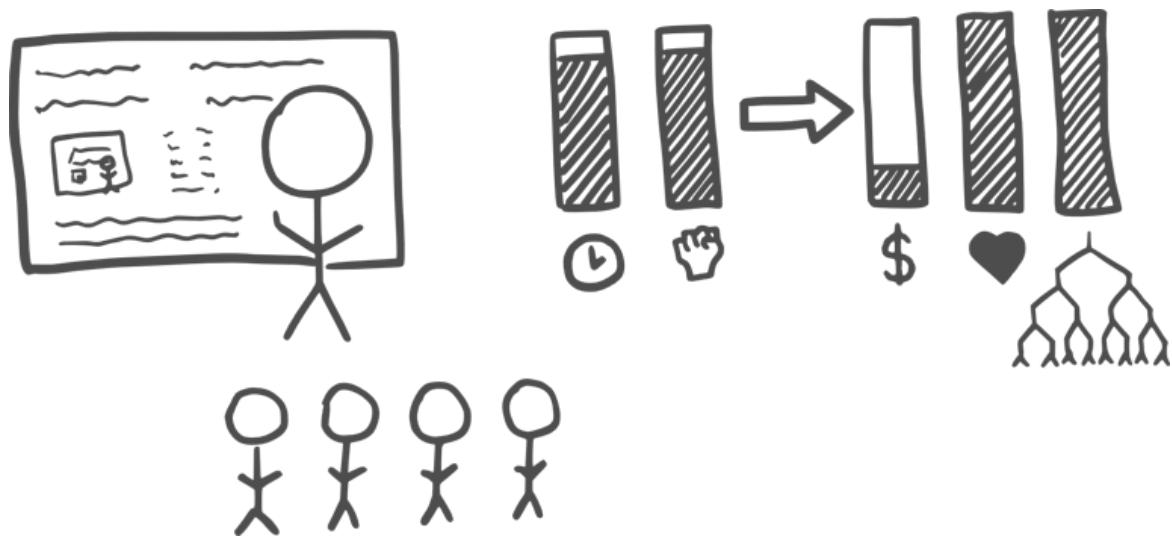
Some people benefit from converting *everything* in their lives to a common currency. You can imagine a particular “chunk” of good, like your favorite menu item at your favorite restaurant, and make productive comparisons. Will this vacation be worth 100 sesame chicken dinner combos? Are you willing to have one less dinner combo per month, to pay for a gym membership? Is the amount of happiness you’re expecting to get out of a frivolous purchase more or less than a couple of dinner combos?



This can be a valuable exercise in many ways, but it's important to recognize that it's a simplification, a shorthand. Most of the things we do involve multiple currencies, and when you try to boil them down to a single number, you'll often find that you're either leaving things out or spending way too much time arriving at exactly the correct appraisal. Your dinner combo may cash out in your head to \$12, but it *also* takes time to order, and saves time otherwise spent cooking. It provides a certain amount of visceral satisfaction, and hits or misses various nutritional goals. It may be part of a weekly tradition with friends where you gain social value. It's complex, with lots of moving parts.

This is true of many currency-type situations. One of the most common exchanges in our society is in the workplace, where we trade time and effort for money. Yet many of us work jobs that pay “less than we’re worth,” because money is not the *only* thing we’re getting out of the transaction—think of altruists and philanthropists working at non-profits, or people taking a risk on their big startup idea. When we consider the value of our time, money is a good first approximation, but it’s rarely the whole picture. If you offer me \$1 per hour to sort pencils, I’ll say no; offer me \$1000 per hour, and I might say yes. By negotiating back and forth, we can get a sense of my default hourly rate for thankless tasks—but that doesn’t

touch on tasks that aren't thankless, or issues of supply and demand and specialization.



Teachers often exchange large amounts of time and effort for moderate amounts of money, but large amounts of personal satisfaction and a chance at greater impact.

**Moral: Costs and values are often made of multiple parts.**

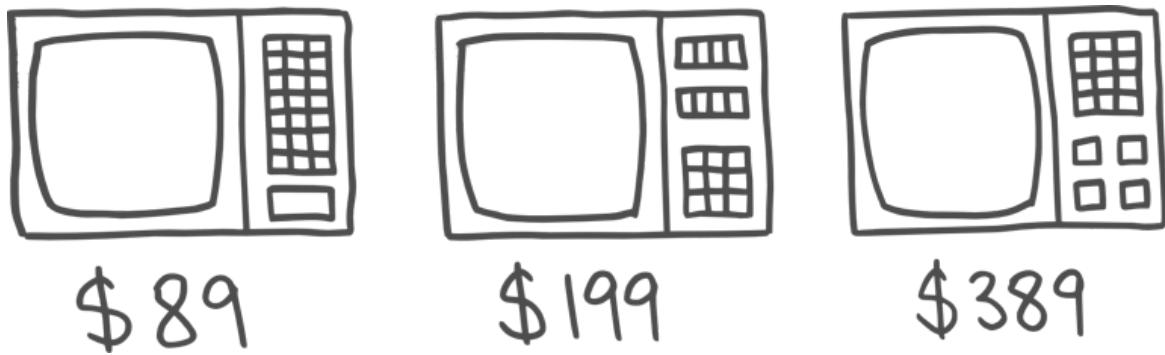
**Moral: Seek simple comparisons, but also mistrust them.**

---

## Part II: Relevant value, relevant cost

Imagine that you're in the market for a new microwave. You're standing in the aisle, looking at three options—one for \$89, one for \$199, and one for \$389. How do you decide?

It may be that you have a certain budget for microwaves, and that's that—sometimes, a particular currency is the overwhelming limiting factor, and if your bank balance is low, all other considerations come second. But imagine that you have room to at least *consider* all three options. What are the relevant details?



Cost is one, obviously. Quality is another—some combination of power, reliability, versatility, and durability. Aesthetics might also be a factor, or energy efficiency, or ease of use.

The consideration that most people miss, in this case, is *time*. A microwave is a device you're likely to use almost every day, perhaps multiple times a day, and a quick Google search shows that the average microwave lasts around nine years. That's somewhere between two and three thousand uses, at a minimum. This means that the difference between a microwave that heats your food properly in two minutes on the first round and one that takes four or five minutes with repeated breaks to check and stir is enormous. It's an extra frustration on or off the pile every day for years; an extra hour saved or wasted every month.

You can do a similar calculation based on how much happiness or satisfaction you get from evenly-heated food versus food that's boiling in some spots and cold in others; the difference between, say, "zero-point-two 'happies' per microwave use" can add up!

$$45 \frac{\text{sec}}{\text{day}} \times 365 \frac{\text{day}}{\text{year}} \times 9 \text{ years} = 40^+ \text{ hrs}$$

$$.2 \frac{\text{happies}}{\text{meal}} \times 300 \frac{\text{meals}}{\text{year}} \times 9 \text{ years} = 540 \text{ happies } (\pm)$$

That doesn't necessarily mean that the most expensive option is always the right choice. But it's a valuable way to reframe the problem. When you're standing in the store, it's easy to think that the *only* tradeoff is between money and quality. It's hard to remember that "quality" has other ramifications, and usually worthwhile to unpack them, at least a little. You

might save a couple hundred bucks on the spot, only to lose a dozen hours in the future—a dozen hours that, for most of us, are ultimately worth much more than the one-time hit to our bank balance.

**Moral: The real cost isn't always on the sticker.**

**Moral: Beware repeated costs—they add up!**

---

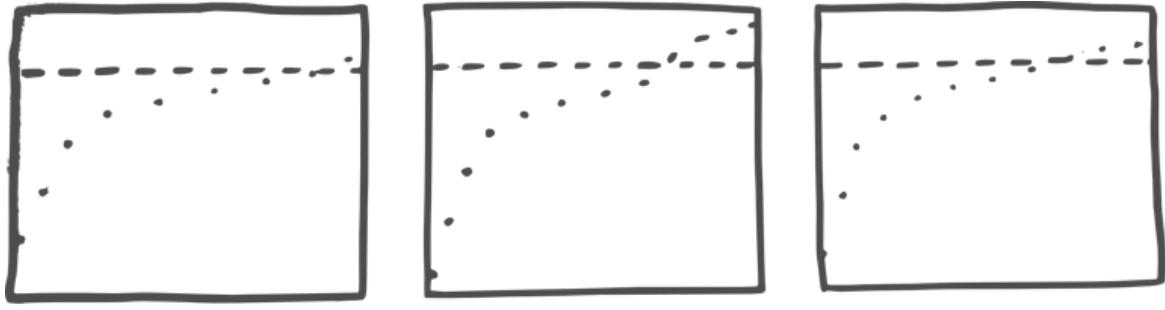
## Part III: Diminishing returns

### SHOPPING TEAMS



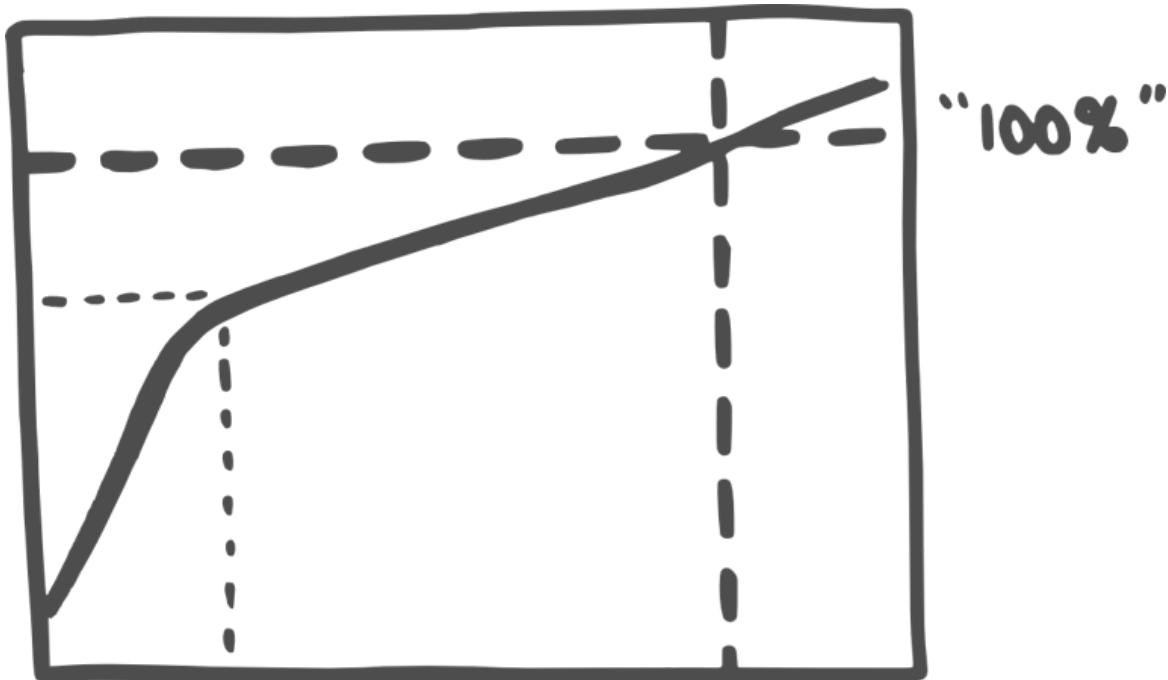
From [xkcd.com](http://xkcd.com)

There is a cost to pursuing any strategy, whether it's in time, money, effort, resources, etc. Most strategies have diminishing returns, meaning that, as you keep at them, you get less and less out of an additional marginal bit of effort. Think about continuing to make sales calls in a small city after you've already tapped all of the obvious buyers, or clicking forward to the tenth page of Google results, or eating your fifth slice of pizza.



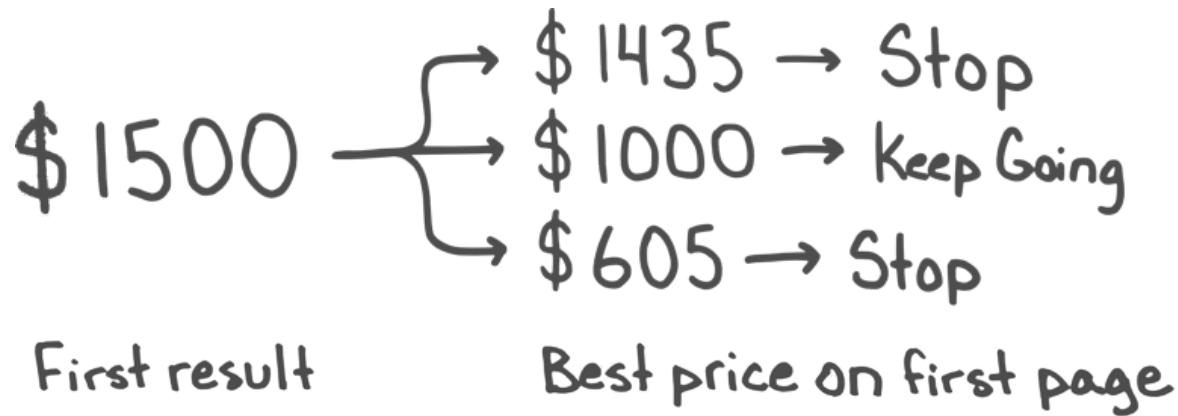
## Running      Coding      Socializing

There's a general principle (often called the Pareto principle or the 80/20 rule) which states that eighty percent of the *results* come from twenty percent of the *effort*. It's not a hard-and-fast rule, of course, and there are situations where it doesn't apply at all. But for many strategies, it's best to put forth strong effort in the early stages, get the bulk of the low-hanging fruit, and then switch to something else. When you start running, or begin coding, or change the way you socialize, you'll see steep improvement that eventually starts to level off.



Consider “information” as an example. The **value of information** is how much better you expect your life to be based on the information you’re seeking—a balance of how much of a difference that information *could* make, and how likely it is that it *will* make that difference. For instance, if you’re searching for plane tickets, more information could conceivably save you hundreds of dollars, but the *odds* of you finding such significant savings may not be clear.

Most of us have an instinctive grasp of this principle. If the first ticket we come across is \$1500, we immediately glance down the page to get a sense of the possible savings—are all of the options in the same range, or do some of them dip down below \$1000? We then spend time and effort accordingly—if it *feels like* an extra ten minutes of digging might save us five hundred dollars, we keep going, and if it feels like we've pretty much seen all there is to see, we stop searching and buy the best ticket we've found so far.



The key is to employ this same metastrategy everywhere it makes sense. Keep your eye on the marginal value of each extra hour, each extra dollar, each extra drop of motivation or discipline, and when that value starts to dip, use that as a reminder to ask yourself: do I expect this strategy to continue paying for itself, or is it time to change course?

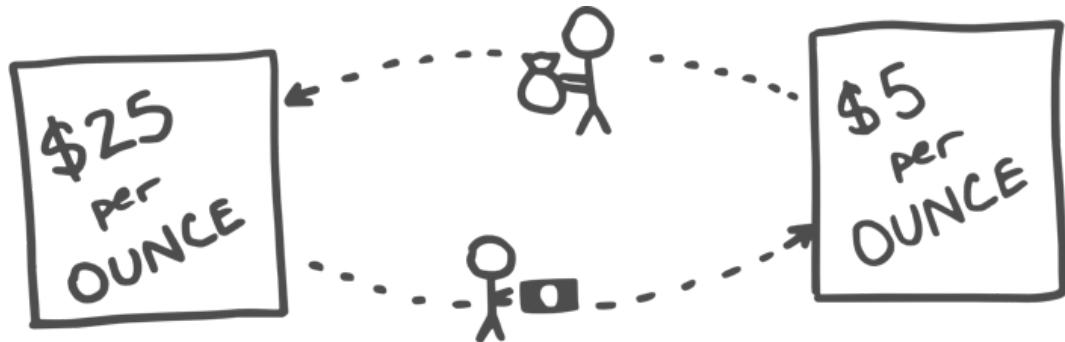
**Moral: Early gains tend to be the largest.**

**Moral: Every strategy eventually stops being worthwhile.**

---

## Part IV: Arbitrage

“Arbitrage” is an economics term that essentially boils down to “take advantage of the fact that things have different prices in different places.” If silver costs \$5 per ounce in one part of the world, and \$25 per ounce in another, and you have the proper logistics in place, you can buy an ounce where it’s cheap, sell it where it’s expensive, then take the \$25 you’ve earned and use it to buy five ounces, sell them both, and so on.



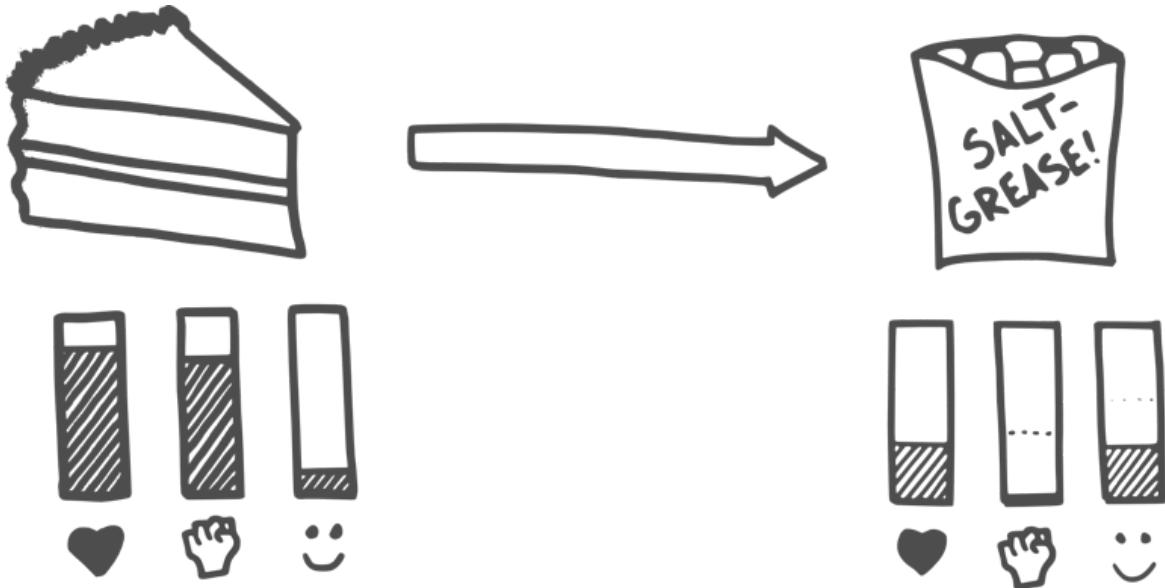
Arbitrage has the effect of leveling out prices—you can’t keep that process going forever, because at some point the supply in the cheap place will drop, and the demand in the expensive place will drop, and things will be *consistent* between the two markets. But in the meantime, you can exploit the inconsistency to make money out of (essentially) nothing.

There are similar opportunities for arbitrage in our own personal “currency markets.” Most of us are inconsistent in how we prioritize time, money, energy, social effort, and other resources—we overspend in some areas and underspend in others, effectively “narbitraging” ourselves. By targeting those inconsistencies and shifting resources around, we can create extra value even without adding anything new to the system.

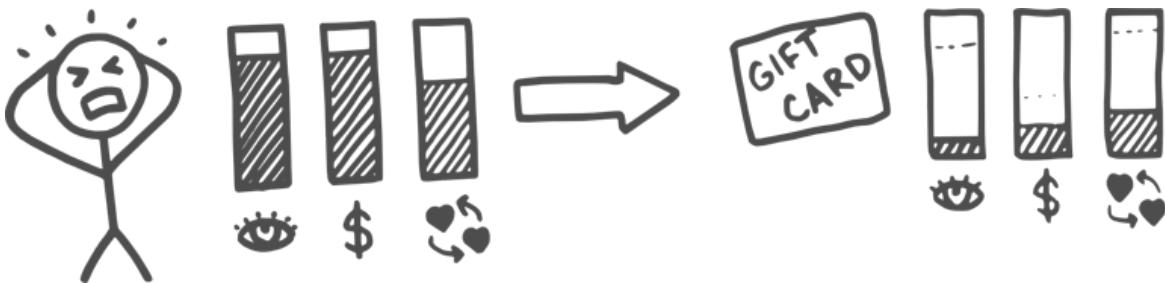
We’ve already touched on one example above—we lean toward buying the cheaper microwave to save money, but may overlook the possibility that buying the more expensive one can save us significant amounts of time, some of which could be used to earn more money than we spent in the first place. Paying attention to the relationship between time and money in the long term changed the calculus of the purchasing decision, likely for the better.

Other examples:

Someone who rigidly holds themselves to a superstrict diet and spends lots of willpower to (e.g.) turn down delicious homemade cake at a party, and then burns out and binges on cheap potato chips three days later. The currencies being traded here are effort, health, and food-related happiness; this person bankrupted themselves on the former, and got compromised versions of both of the latter. If instead they had eaten some cake, they would have retained willpower, taken a comparable hit to their health, and gotten significantly more food-related happiness—a better outcome.



Someone who consistently struggles to come up with thoughtful, meaningful gifts for their family and friends, and who usually spends the week before Christmas or birthdays wracked with guilt and stressing out over what to make or buy. The currencies being traded here are time, attention, and goodwill/warm fuzzies; this person spends a large amount of both of the former for uncertain results on the latter. If they instead create a single, easy place to store gift ideas year-round, they can decrease the costs in time and attention and be more likely to pinpoint and remember the things their families and friends actually want.



Someone who spends dinner time bouncing back and forth between work texts/emails and conversation with family and roommates. Currencies being traded here are time, attention, effectiveness, and the other diners' sense- that-they-matter; by juggling two important things, this person is likely to fail at both. If they instead shorten their dinner commitment to twenty minutes, but are *fully present* before going back to work, they can spend the same amount of time, reduce attention overhead and switching costs, and improve both their ability to get the work done and to show affection for their family and roommates.

The most important piece of this puzzle is the recognition that **attention to tradeoffs ≠ being cold and calculating**. Often we feel a little strange doing things like arranging to see all of our friends at the same party, because there is a sense that this is cheating or manipulative or somehow disingenuous. And it's true that making small, specific sacrifices in the process of seeking arbitrage can *draw your attention* to tradeoffs that are somewhat uncomfortable.

But it's important to recognize that *those tradeoffs were already happening*. It's like hospital administrators making tough calls between expensive procedures for sick children and new equipment or raises for surgeons. We already trade time against money against effort against happiness against social capital—we can do so blindly, and hope for the best, or we can think about them carefully and deliberately, and take advantage of opportunities to get more of *everything*. If your schedule is overloaded, you're already shortchanging your friends by being distracted or exhausted or otherwise sort-of-not-really-there for them; rearranging things to see more of them in groups isn't taking anything from them, and it's *giving back* to yourself.

(And if it turns out that it *is* taking something from them—if you discover that some of those relationships need more one-on-one time than you thought—you can change the plan again!)

This is the key. You have limited amounts of time/money/effort/etc., so it makes sense to waste them as little as possible—you're not looking to sacrifice one part of your life for the sake of another, you're looking for ways to increase one part at no cost to the other, or to raise the overall available amount of every currency by fixing the leaks.

**Moral: Identify all relevant currencies, and note which are being spent faster or are more valuable**

**Moral: Proper arbitrage isn't win-lose, it's win-win. You can reinvest the recouped resources however you want, including right back into the thing you just made more efficient.**

---

## Part V: Opportunities for growth

The following are some areas where many CFAR alumni have found significant opportunities for improving the tradeoffs they were making:

- Rearranging commutes or other regular time commitments
  - Improving reading or typing speed; switching to audio books
  - Using earplugs, eye masks, and white noise to improve sleep quality
  - Regular re-evaluations of job, career, salary, project, team role, etc.
  - Efficiency systems like keyboard shortcuts, email routines, & to-do lists
  - “Batching” small recurring tasks to avoid switching costs
  - Making one-time purchases (including “purchases” of time, energy, or social effort) that remove or reduce the cost of a repeated expense
  - Using Craigslist, Uber/Lyft, Ebay, OKCupid/match.com, mailing lists, and event calendars
- 

## Units of Exchange—Further Resources

Explicit calculations are useful in part because people’s intuitions often have a hard time dealing with quantities (for a review, see Kahneman, 2003). In a classic study on scope neglect, people were willing to spend about as much to save 2,000 birds as to save 200,000 birds. Similar insensitivity to variations in quantity, which Kahneman (2003) calls “extension neglect,” arise in other contexts. For example, people’s evaluations of an experience (such as a medical procedure without anesthesia) tend to be relatively insensitive to its duration (compared to the peak level of emotion).

Kahneman, D. (2003). *A perspective on judgment and choice: Mapping bounded rationality*. American Psychologist, 58, 697-720. <http://tinyurl.com/kahneman2003>

---

People are more sensitive to quantities when they can make side-by-side comparisons of multiple options which vary on that quantity or when they have enough familiarity with the subject matter to have an intuitive sense of scale, but in the absence of these conditions a person’s intuitions may be essentially blind to the magnitude of the quantity (Hsee, 2000). In order to incorporate the magnitude in one’s judgment, it may be necessary to engage in explicit effort to make sense of it.

A review article on “attribute evaluability,” which is the extent to which a person is sensitive to quantitative variations in an attribute:

Hsee, C. K. (2000). *Attribute evaluability and its implications for joint- separate evaluation reversals and beyond*. In D. Kahneman & A.Tversky (eds.), Choices, Values and Frames. Cambridge University Press. <http://goo.gl/3IXoD>

---

Research on decision making suggests that people who care a lot about making the best decision often neglect the implicit costs of the decision making process such as time and money. For example, they might spend a lot of time trying to pick a good movie to watch (neglecting the time cost) or channel surf while watching television (neglecting how dividing attention can reduce enjoyment); self-reports of both behaviors have been found to correlate with personality trait of “maximizing.”

Schwartz, B., Ward, A., Monterosso, J., Lyubomirsky, S., White, K., & Lehman, D.R. (2002). *Maximizing versus satisficing: Happiness is a matter of choice*. Journal of Personality and

Social Psychology, 83, 1178-1197. <http://goo.gl/HlImnQ>

Alumni Lincoln Quirk's essay on how to put a dollar value on one's time: <http://goo.gl/fVDuFj>

---

A blog post with several vignettes in which VOI calculations are relevant:

[http://lesswrong.com/lw/85x/value\\_of\\_information\\_four\\_examples/](http://lesswrong.com/lw/85x/value_of_information_four_examples/)

---

The introduction to Aaron Santos's book provides a simple guide for how to make rough estimates of quantities, and how to break a difficult-to-estimate quantity into components. The rest of the book contains sample problems for practicing Fermi estimation.

Santos, Aaron (2009). *How Many Licks? Or, How to Estimate Damn Near Anything.* <http://goo.gl/8ytNye>

---

Related comics and essays by Randall Munroe:

"[Is It Worth the Time?](#)"

"[Paint the Earth](#)"

"[A Mole of Moles](#)"

# In defense of flailing, with foreword by Bill Burr

I don't give a [expletive] who you are, if the world is ending, and you're getting chased by zombies, you're not running around saying "oh golly gee. Oh heck... Aw jimmety cricket!"... You know?... It's the *end of the world with zombies*. From the beginning, once they discover the zombies, to the end of that movie where they hopefully solve the problem, should be a bunch of people - no wait - 85% of the people going "[array of sorted expletives] WHAT ARE WE GONNA DO OH MY GOD" and then the other 15% should be grabbing them by the shoulders going "FOR CHRIST'S SAKE GET A [expletive] HOLD OF YOURSELF!"... That should be most of the dialogue in that movie.

- [Comedy man on existential risk](#)

CFAR has a notion of "flailing". Alone on a desert island, if you injure yourself, you'll probably think fast about how to solve the problem. Whereas injuring yourself around friends, you're more likely to "flail": you'll lean into things that demonstrate your pain/trouble to others.

Flailing can in fact be counterproductive. I will admit that if you're a MIRI engineer, and you start to flail in front of the other MIRI engineers, that is probably in poor taste. Everyone at MIRI already *knows* you are in trouble and is doing their best. Many of them are as scared as you are, and are trying to suppress their own instinct to flail.

But, just like crying on the floor, flailing does in fact serve legitimate personal and social purposes. People evolved shared abilities to signal distress *for a reason*. Before March of this year, on the rare occasion I saw the Utterly Horrifying Situation explicitly acknowledged, the guy on my laptop screen devoting his life to the problem would always follow by saying: "Don't Panic. Do not alert the other people next to you that you might be panicking if you are. Don't honestly convey how worried you are about the problem. Deliver your argument in reasoned, unemotional tones, like you are considering a hypothetical, not like you are explaining an asteroid is headed towards earth. If you must break down, do it quietly and in a separate room, because otherwise people might not think you're Rational and by extension your community is not Rational."

We are not on a desert island. A large fraction of the basic coordination problem stems from the complete absence of public and institutional understanding of how serious the situation has gotten. It clearly wouldn't solve everything, but a world in which most people understand that sans serious coordination it's going to end seems closer to organizing a solution than one where almost nobody does. And if you're a concerned citizen trying to alert others, you should realize there's a distinct problem with suppressing your emotions.

That problem is: people hear philosophical-sounding arguments as to why we might be doing something very bad all the time. Normies take a brief look at these arguments, and then check out the emotional state and behavior of the person delivering them. The way you act **and** speak telegraphs to the other person how to react to the problem and the degree to which you yourself care about it.

If you are a programmer at FaceGoog who really does sound like they read something very scary on the internet, but puts 40% of their income into their 401k, bystanders who know these facts about you will probably not do anything about your pet doomsday scenario either, no matter how solid your reasoning is, because you are not behaving how they'd expect someone concerned about doomsday to behave. If you're an AI safety researcher who has a job working on the problem but you explain it to others like you're reading a weather forecast, they will probably be even less inclined to believe you, because people in the Real World will assume you should have an emotional attachment to the significance of your pet cause and have had *lots* of time to justify it to yourself. "This just sounds like a LARP, why aren't you doing anything about it then" is by far the most common unmanageable reaction I get after an hour conversation trying to raise the alarm for people I know, and I have little to no good explanation except "I dunno, maybe I'm a terrible/illogical person".

In fact, since the "[MIRI announces new "Death With Dignity" strategy](#)" post, I've had a bit of time to reflect on why I've been working on a miscellaneous startup for years, even though I would've told anybody who asked me that the most likely outcome for myself is death before 30 years old, and why I suddenly now feel like I have to wake the fuck up. The reason is I've had these misgivings is probably because *Eliezer started flailing*. I don't think "Death with Dignity" is actually a helpful way, psychologically, of looking at the problem, but the post highlighted Eliezer's internal world model to me in a second-order way that flipped a switch. It was the *way he communicated*, how his tone was consistent with what my hindbrain expects from the alarmed, *not* the content, that started to get me to think more seriously about what on Earth I was trying to do with my life. Same thing with the [AGI Ruin](#) post. As Zvi puts it: "One could also propose making it not full of rants, but I *don't* think that would be an improvement. The rants are *important*. The rants contain *data*. They reveal Eliezer's cognitive state and his assessment of the state of play. *Not* ranting would leave important bits out and give a *meaningfully misleading impression*."

There's probably a line beyond which panicking or showing too much emotion about how the world is ending makes you look crazy, and where that line is placed is certainly different depending on how well you personally know whoever you're talking to. However, for this particular issue it seems very rare, in practice, for people to cross that line. On the contrary, it seems like most people pretend to be calm or evade the actual subject of pending doom to a point that completely deflates their attempt at mobilizing or convincing others. Eliezer@pre-2022 will give a completely dry explanation of the problem or the factors surrounding the problem, and no matter how safely inside the Overton window he tries to be, anybody who dislikes him will just make up some psychoanalytic nonsense about how he's doing it because he wants to validate the importance of his IQ. So unless you sound incoherent to the people who were going to believe you anyways, I think you should be honestly expressing how you feel and the degree to which the problem concerns you. If you believe with 90% probability everyone is going to die in the next fifteen years, *no one seems to understand that* after talking with you about the problem, and it's not your deliberate intention to hide your beliefs, you're not being explicit enough.

# Air Conditioner Repair

A thing happened to me, seemed worth sharing.

## What Happened

A few weeks ago, the in-wall air conditioner in our bedroom ceased to function.

Troy Barnes was unavailable, so instead we talked to the super, he recommended a company called Amhac, we ignored normative determinism (e.g. "Am a hack") and called them. They charged a rather large amount simply to come out and take a look, with a diagnostic fee and also a noticeably generous per-hour fee.

Then, a few days before they were supposed to show up, our *other* in-wall air conditioner sprung a leak – or rather, the leak got big enough that we noticed it – and the super advised us we had to shut it down until it was fixed, but he thought that one would be a simple fix plus the addition of a failsafe that should have been there to stop the leak; he wasn't sure about the bedroom one.

The repairmen showed up, but they hadn't completed the necessary certificate of insurance to be let into the building – this policy is even more annoying and friction generating than it sounds – and we had to reschedule while they sorted this out. A number of rather hot nights later, they managed to come in.

Their report was that both air conditioners were unfixable. The one in the bedroom was completely shot. The one in the living area was fixable in theory, they said, but due to some EPA regulation it wasn't possible to fix it.

They could send over a contract for the new units, it would be... \$28,000.

I Googled for various new air conditioners, and couldn't find ones that would be that expensive. The guy explained I would need two of each unit, for four units total.

I asked some trusted friends about all this, as well as the super, and all agreed a second quote would be a very good idea. I asked the super who else had done work in the building, and he remembered M.LaPenna Refrigeration, Inc. They charged by the hour as well, but without a 'diagnostic fee' up front.

Which on reflection should have been a hint about the first company. If you're charging for the labor, as you do, starting with minute one, why is there an additional diagnostic fee? It doesn't actually make any sense.

This time, they took care of the certificate of insurance as fast as the insurance company could handle it, then got a guy out to look the same day that got approved. Almost as if things were urgent. Different attitude.

Instead of two workers, this time there was one, who was keen to explain to me what was going on as he worked.

The super was busy, so I explained the situation as best I could and he got to poking around in various places.

Two hours later, both air conditioners were working again, and they sent out for the part to install the failsafe.

I asked the man if there was any way the other repairmen could have made an honest mistake saying the units needed to be replaced. His answer: "No."

If I hadn't checked, I wonder how much they finally would have tried to get me for, but I'm sure it was a lot.

## What To Do About That First Company?

I'm not sure. I emailed my contact to say what happened and suggested a full refund would be appropriate. The response was that the person was on vacation for weeks (with no warning). Which did not endear me.

If it was one unit, I could imagine an honest mistake. But it wasn't. It was *two* units, with distinct setups, experiencing distinct problems. This was not a mistake.

What am I supposed to do now? Chargeback? Report to better business bureau? Report to someone else? Do something else? What's the responsible thing to do here?

I don't know.

## Thoughts and Takeaways

What about takeaways in general?

First of all, *get a second opinion*. Do not trust contractors of any kind, who you don't have damn good reason to trust, who tell you that you need something massively expensive or how much that something should cost until it has been verified. However much the cost in delay, mild social awkwardness and an extra payment for the double check, not double checking is malpractice.

Quality is highly variable. Some people are great. Some people are less great. Others are [out to get you](#). I had the same experience when I explored getting a new wall, with proposals differing in cost by an order of magnitude. The person someone recommended did not listen at all, then when I told him what he was proposing was not at all what I'd asked, responded with 'well if we're going to go back and forth then you need to pay for the proposal.' We found a much better option, but ended up deciding what we had was fine.

Second, *remember to be scope sensitive and give proper attention when there are bigger stakes*. A small number of relatively big decisions are worth quite a lot, yet there will be that temptation to be done with it to avoid the stress and the mild social awkwardness. Resist this.

Third, *air conditioner repair seems like a damn fine business*. This was most certainly truth in television. The repair role isn't anything in the job that an average person couldn't learn, after which you're making three figures an hour while doing an actual physical useful thing. Centrally, you solve puzzles, figure out what's wrong and how to fix it, and make things work again. Very not alienating.

It sure seems like it beats a lot of ‘white collar’ jobs I’ve seen, and it’s open to pretty much anyone. If you run the business yourself, that seems even better. There are some barriers to entry there, especially starting capital, but again it seems pretty sweet, and you do well by doing good.

It’s also a job that seems relatively safe from automation in the medium term.

In general, the category of ‘physical work to make physical things work that requires skills but which can be learned’ seems like it pays pretty well and has strong demand.

Don’t get me wrong. It’s not especially high on my list of things I would try doing, but it seems worth putting on the list of pretty damn good options.

Fourth, I suppose I should get that maintenance contract up and running?

# Pivotal outcomes and pivotal processes

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

**tl;dr:** If you think humanity is on a dangerous path, and needs to "pivot" toward a different future in order to achieve safety, consider how such a pivot could be achieved by multiple acts across multiple persons and institutions, rather than a single act. Engaging more actors in the process is more costly in terms of coordination, but in the end may be a more practicable social process involving less extreme risk-taking than a single "pivotal act".

**Preceded by:** ["Pivotal Act" Intentions: Negative Consequences and Fallacious Arguments](#)

[This post is also available on the [EA Forum](#).]

In the preceding post, I argued for the negative consequences of the *intention* to carry out a pivotal act, i.e., a single, large world-changing act sufficient to 'pivot' humanity off of a dangerous path onto a safer one. In short, there are negative side effects of being the sort of institution aiming or willing to carry out a pivotal act, and those negative side effects alone might outweigh the benefit of the act, or prevent the act from even happening.

In this post, I argue that it's still a good idea for humanity-as-a-whole to make a large / pivotal change in its developmental trajectory in order to become safer. In other words, my main concern is not with the "pivot", but with trying to get the whole "pivot" from a single "act", i.e., from a single agent-like entity, such a single human person, institution, or AI system.

## Pivotal outcomes and processes

To contrast with pivotal acts, here's a simplified example of a *pivotal outcome* that one could imagine making a big positive difference to humanity's future, which in principle could be brought about by a multiplicity of actors:

- (**the "AI immune system"**) The whole internet — including space satellites and the internet-of-things — becomes way more secure, and includes a distributed network of non-nuclear electromagnetic pulse emitters that will physically shut down any tech infrastructure appearing to be running rogue AI agents.

(For now, let's set aside debate about whether this outcome on its own would be pivotal, in the sense of pivoting humanity onto a safe developmental trajectory... it needs a lot more details and improvements to be adequate for that! My goal in this post is to focus on how the outcome comes about. So for the sake of argument I'm asking to take the "pivotality" of the outcome for granted.)

If a single institution imposed the construction of such an AI immune system on its own, that would constitute a *pivotal act*. But if a distributed network of several states

and companies separately instituted different parts of the change — say, designing and building the EMP emitters, installing them in various jurisdictions, etc. — then I'd call that a *pivotal distributed process*, or *pivotal process* for short.

In summary, a pivotal outcome can be achieved through a pivotal (distributed) process without a single pivotal act being carried out by any one institution. Of course, the "can" there is very difficult, and involves solving a ton of coordination problems that I'm not saying humanity will succeed in solving. However, aiming for a pivotal outcome via a pivotal distributed process definitively seems safer to me, in terms of the dynamics it would create between labs and militaries, compared to a single lab planning to do it all on their own.

## Revisiting the consequences of pivotal act intentions

In [AGI Ruin](#), Eliezer writes the following, I believe correctly:

- *The reason why nobody in this community has successfully named a 'pivotal weak act' where you do something weak enough with an AGI to be passively safe, but powerful enough to prevent any other AGI from destroying the world a year later - and yet also we can't just go do that right now and need to wait on AI - is that nothing like that exists. There's no reason why it should exist. There is not some elaborate clever reason why it exists but nobody can see it. It takes a lot of power to do something to the current world that prevents any other AGI from coming into existence; nothing which can do that is passively safe in virtue of its weakness.*

I think the above realization is important. The un-safety of trying to get a single locus of action to bring about a pivotal outcome all on its own is important, and it pretty much covers my rationale for why we (humanity) shouldn't advocate for unilateral actors doing that sort of thing.

Less convincingly-to-me, Eliezer then goes on to (seemingly) advocate for using AI to carry out a pivotal act, which he acknowledges would be quite a forceful intervention on the world:

- *If you can't solve the problem right now (which you can't, because you're opposed to other actors who don't want [it] to be solved and those actors are on roughly the same level as you) then you are resorting to some cognitive system that can do things you could not figure out how to do yourself, that you were not close to figuring out because you are not close to being able to, for example, burn all GPUs. Burning all GPUs would actually stop Facebook AI Research from destroying the world six months later; weaksauce Overton-abiding stuff about 'improving public epistemology by setting GPT-4 loose on Twitter to provide scientifically literate arguments about everything' will be cool but will not actually prevent Facebook AI Research from destroying the world six months later, or some eager open-source collaborative from destroying the world a year later if you manage to stop FAIR specifically. **There are no pivotal weak acts.***

I'm not entirely sure if the above is meant to advocate for AGI development teams planning to use their future AGI to burn other people's GPU's, but it could certainly be read that way, and my counterargument to that reading has already been written, in

["Pivotal Act" Intentions: Negative Consequences and Fallacious Arguments](#). Basically, a lab X with the intention to burn all the world's GPUs will create a lot of fear that lab X is going to do something drastic that ends up destroying the world by mistake, which in particular drives up the fear and desperation of other AI labs to "get there first" to pull off their own version of a pivotal act. Plus, it requires populating the AGI lab with people willing to do some pretty drastically invasive things to other companies, in particular violating private property laws and state boundaries. From the perspective of a tech CEO, it's quite unnerving to employ and empower AGI developers who are willing to do that sort of thing. You'd have to wonder if they're going to slip out with a thumb drive to try deploying an AGI against *you*, because they have their own notion of the greater good that they're willing to violate *your* boundaries to achieve.

So, thankfully-according-to-me, no currently-successful AGI labs are oriented on carrying out pivotal acts, at least not all on their own.

## **Back to pivotal outcomes**

Again, my critique of pivotal acts is not meant to imply that humanity has to give up on pivotal *outcomes*. Granted, it's usually harder to get an outcome through a distributed process spanning many actors, but in the case of a pivotal outcome for humanity, I argue that:

1. it's *safer* to aim for a pivotal outcome to be carried out by a distributed process spanning multiple institutions and states, because the process can happen in a piecemeal fashion that doesn't change the whole world at once, and
2. it's *easier* as well, because
  1. you won't be constantly setting off alarm bells of the form "Those people are going to try to unilaterally change the whole world in a drastic way", and
  2. you won't be trying to populate a lab with AGI developers who, in John Wentworth's terms, think like "villains" ([source](#)).

I'm not arguing that we (humanity) are going to *succeed* in achieving a pivotal outcome through a distributed process; only that it's a safer and more practical endeavor than aiming for a single pivotal act from a single institution.

# Air Conditioner Test Results & Discussion

Background is in the [preregistration post](#). This post will assume you've read that one.

First, the headline result: my main prediction was wrong. Paul's predictions were also wrong. By the preregistered metric, the second hose improves performance by much less than either of us expected. That said, it looks like the experiment's main metric mostly did not measure the thing it was intended to measure, which is why we were both so far off.

I did collect a bunch of data, which allows us to estimate the air conditioner's performance in other ways. That analysis was not pre-registered and you should therefore be suspicious of it, but I nonetheless believe that it gives a more accurate view of the air conditioner's performance. Main takeaway of that analysis is that the air conditioner performs about twice as well with two hoses as with one. Going by that analysis, both Paul's formula and my prediction were correct.

## Experiment Setup

Here's the air conditioner with the cardboard "second hose" attached:

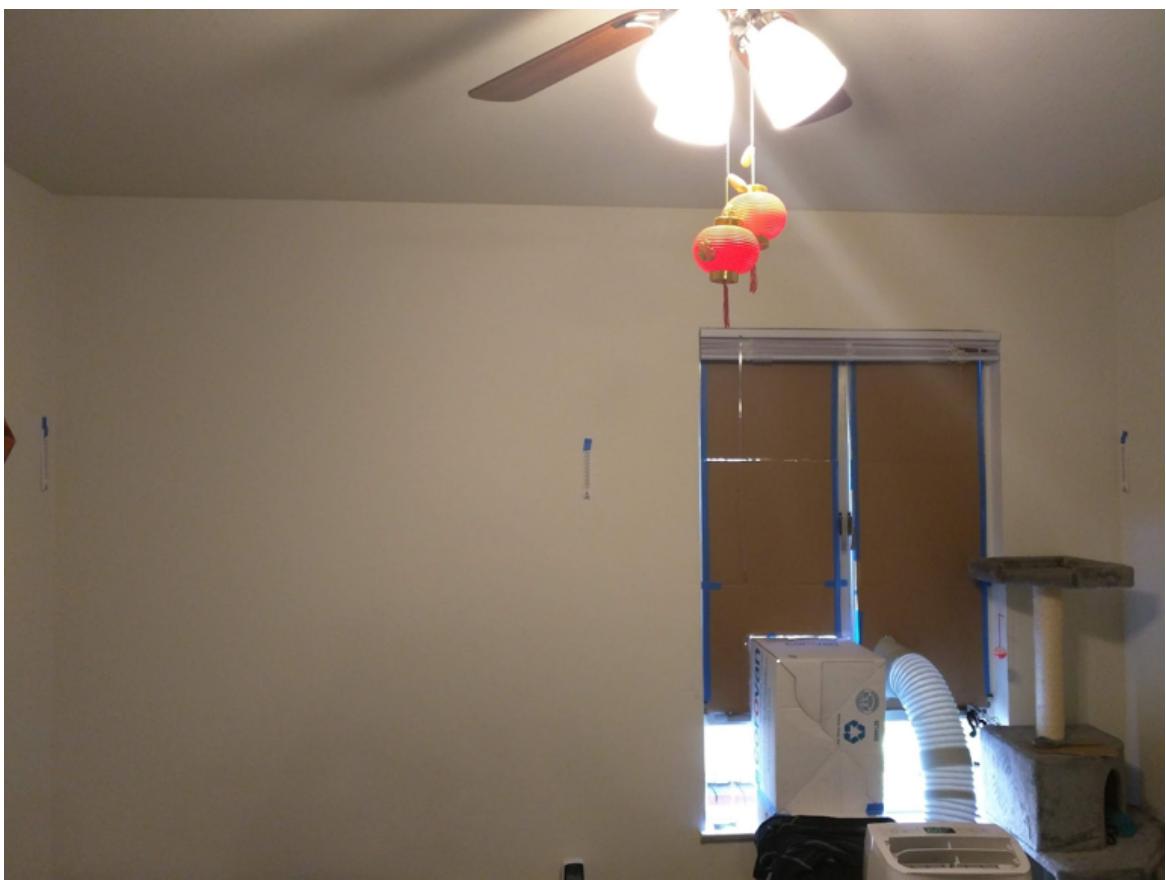


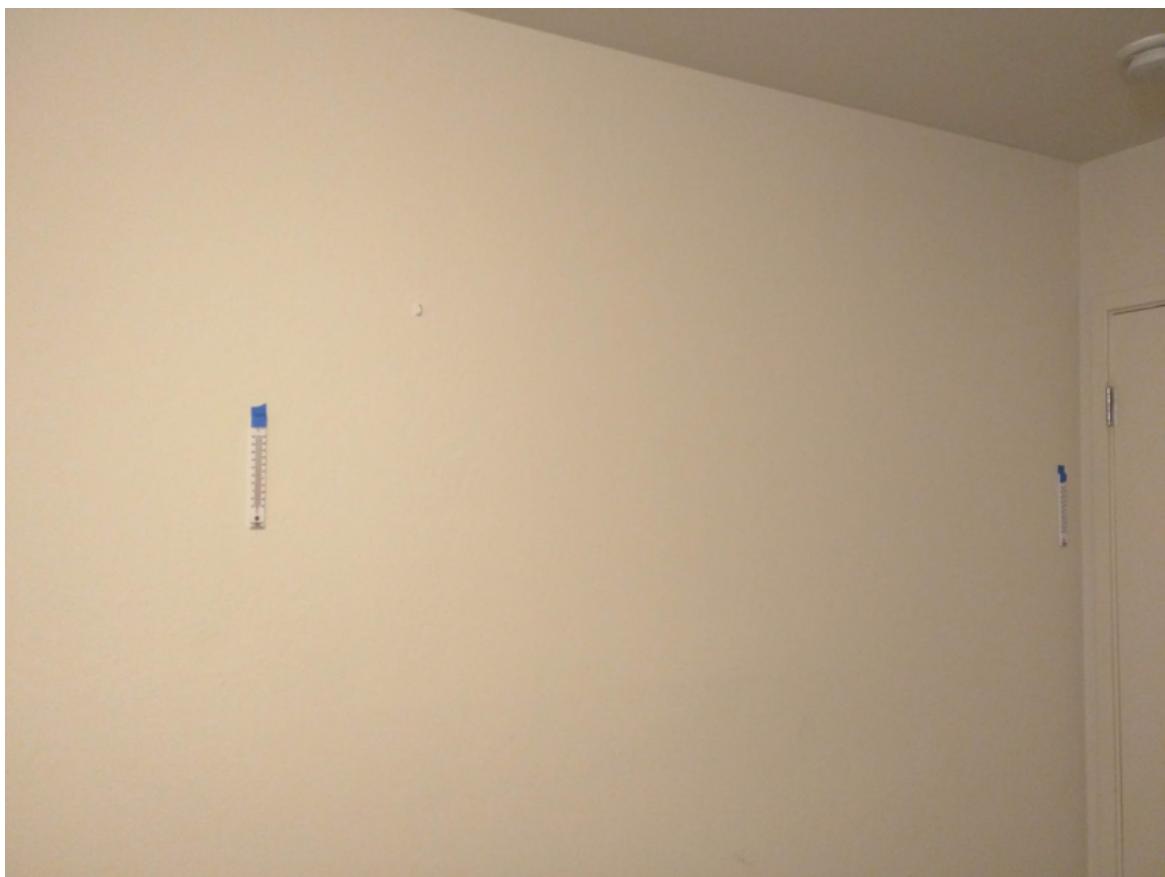


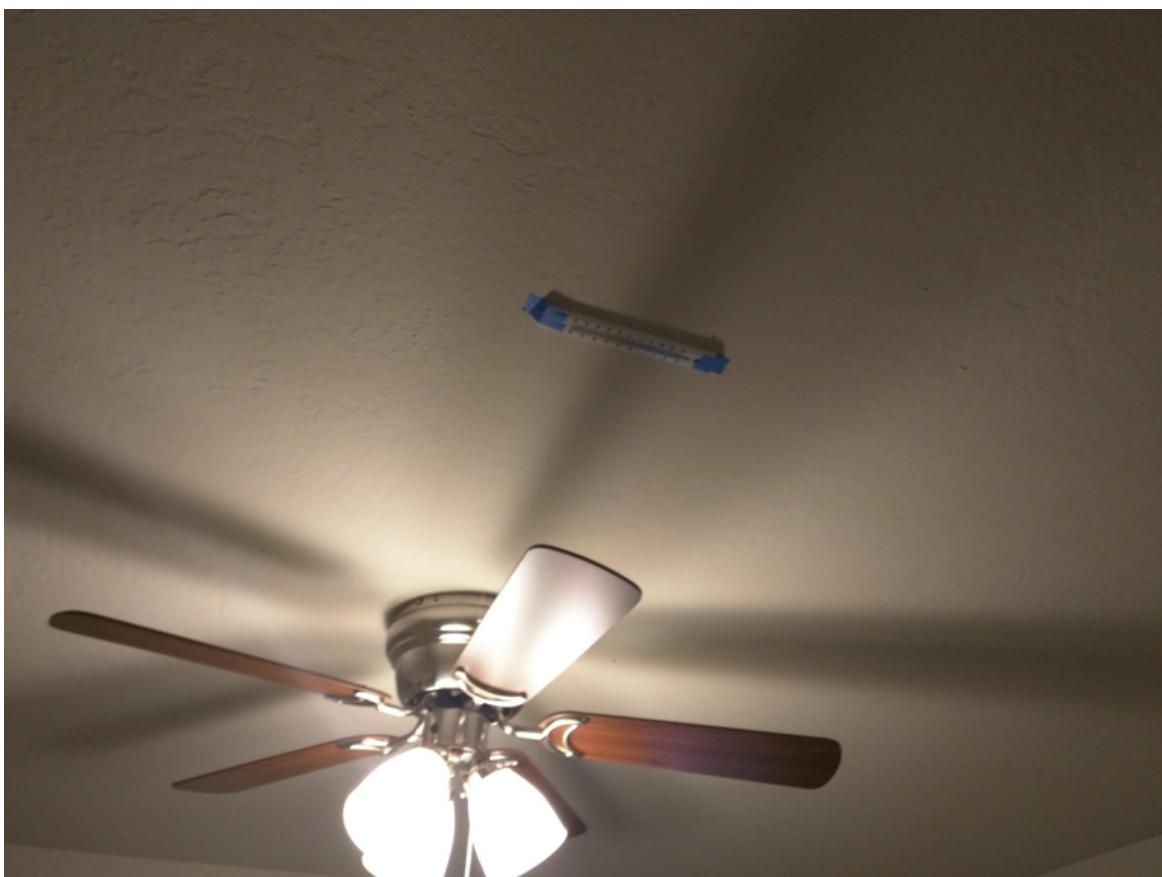
The t-shirts are to cover gaps. It doesn't actually need to be airtight.

I hung thermometers at each corner of the room, in the middle of each wall, and on the ceiling in the center of the room. I also placed one thermometer in the inlet (i.e. the hole in

the window covering through which the cardboard hose draws air), the outlet (i.e. the hole in the window covering through which the other hose blows air), and outside on the balcony.









I ran a few different tests:

- Air conditioner with and without the cardboard intake hose, on low fan
- Air conditioner with and without the cardboard intake hose, on high fan

- Air conditioner off (control)

For both the 1-hose and control tests, I left the inlet hole in the window covering open. (This was mainly to reduce infiltration from places other than outside.)

Assorted Notes:

- I did try the experiment previously (about a month ago), and ran into issues which resulted in some minor changes to the experiment setup; info about that is [here](#).
- None of the experiment setup, thermometers, or the room were in direct sun.
- The time for each test was mostly determined by when I had meetings scheduled, and when the temperatures seemed to stop changing.

## Results

Data is [here](#). Some numbers:

- Outdoor temperature was 85-88°F (29.4 - 31.1°C) for most of the testing, though it dropped to 82°F (27.8°C) in the evening during the control test
- Temperature difference between outdoor and indoor (higher is better), in each test:
  - Low fan: 20.6°F (11.4°C) with one hose, 22.7°F (12.6°C) with two hoses
  - High fan: 18.7°F (10.4°C) with one hose, 22.2°F (12.3°C) with two hoses
  - Control: 13.1°F (7.3°C)
- Temperature variance across the room was fairly high (~2.5 - 3.0°F, or ~1.4 - 1.7°C), and consistent (i.e. measurements 30 minutes apart had similar relative temperature patterns)
- Exhaust temperature was 98 - 100°F (36.7 - 37.8°C) with one hose, 112 - 119°F (44.4 - 48.3°C) with two

## Analysis

My main prediction was that the outdoor-indoor temperature difference would be at least 50% greater with two hoses than with one hose, with my median estimate around 100% (i.e. a factor-of-two difference). That was definitely wrong: the actual number was 10% with fan on low, 19% with fan on high.

Paul's prediction for the same number [was](#) 33-43%, though that was based on some very rough guesses about indoor, outdoor and exhaust temperatures. Using Paul's formula with the actual indoor, outdoor and exhaust temperatures from the tests gives predictions anywhere from 75% to 180%, consistent with my own factor-of-two median guess. (The difference is mainly because the exhaust temperature was considerably lower than Paul had speculated. It really is a very shitty air conditioner.)

So experimentally, the difference came out way lower than anybody guessed. Why?

The result from the control test basically tells us the answer. With the air conditioner off, the room did not return to anywhere near outdoor temperature over the course of an hour. That implies some combination of:

- Very slow equilibration, such that the other test results were probably also not near steady-state.
- Indoor temperatures driven more by (cooler) temperatures in neighboring rooms, rather than outdoor temperature.

The high and consistent-over-time variance of temperatures in different locations within the room also points toward some combination of these two issues. I would guess the second

issue is more important than the first, though I'm not confident in that.

One relatively simple way to correct for the problem: use the temperature from the control test as the baseline temperature, rather than using the outdoor temperature as baseline. If we do that, then with the fan on high the results are:

- AC cools the room by 2.6°F (1.4°C) relative to control with one hose
- AC cools the room by 5.1°F (2.8°C) relative to control with two hoses (despite the outdoor temperature being slightly higher during the two-hose test)

Two hose performs better by about a factor of two, consistent with both Paul's formula (using the real indoor/outdoor/exhaust temperatures) and my guess.

## My Updates

### On the Experiment

First and foremost: I made a confident wrong prediction, so there better be *some* substantial update from that. My main update from that is to put even more weight on "the specific metric you choose will inevitably fail to measure the thing you thought it would measure, especially on your first try". That's something I already knew on some level, I already considered it the main bottleneck to making prediction markets actually useful in practice, and in hindsight it's embarrassing that I didn't put more weight on it when making the air conditioner prediction.

Second: I've basically thrown out my pre-registration and done a bunch of analysis which disagrees with the preregistered analysis. I do think that was the right choice for maximal epistemic accuracy (and in fact is very often the right choice for maximal epistemic accuracy, because the specific metric you choose will inevitably fail to measure the thing you thought it would measure, especially on your first try). But it's important to increment the counter in the back of one's head when doing that, and occasionally go re-examine to see if one is systematically steering away from some undesired conclusion. Counter incremented.

Third: shortly after I put up the preregistration post, both [Paul](#) and [ADifferentAnonymous](#) left comments explaining where Paul's formula came from, and I updated a lot toward that formula being a good model based on the thermodynamic argument they gave. (The main thing the formula leaves out is extra waste heat generated with one hose vs two, and I still had some uncertainty about that.) After seeing the (admittedly very rough) agreement between the formula and the performance-relative-to-control, I currently think the formula is basically correct.

Fourth: both the temperature-change-relative-to-control and the rough thermodynamic calculation (i.e. Paul's formula) with real exhaust temperature point to two hose performing about twice as well as one. I think that's probably about right (modulo some large error bars, of course, since none of this was super precise).

### On One vs Two Hose

Now for the real claim of interest: that single hose air conditioners are "stupidly inefficient in a way which I do not think consumers would plausibly choose over the relatively-low cost of a second hose if they recognized the problems".

That claim boils down to a cost-benefit analysis, and this whole experiment has been about the benefit. What about the cost side? Air conditioner hoses similar to the one my unit uses [cost about \\$20 on amazon](#). Presumably the actual cost-to-the-manufacturer is lower, especially if they're shipping in a box with the rest of the air conditioner. So, the second hose

adds at most \$20 to a \$300-500 air conditioner. It also adds a little bit more annoying fiddliness when setting up the AC.

That cost sounds like it is very obviously worth paying for a factor-of-two performance improvement. It would be very obviously worth paying for the second hose even if my estimates of performance improvement are quite far off; the performance improvement would have to be well below 30% before I'd even start to consider a second hose not-obviously-worthwhile. The marginal cost is just so small.

So, yeah, I do think that single hose air conditioners are stupidly inefficient in a way which I do not think consumers would plausibly choose over the relatively-low cost of a second hose if they recognized the problems.

## On Civilizational Adequacy

The air conditioner was intended as an example in which a product is shitty in ways the large majority of consumers don't notice, and therefore market pressures don't fix it. Two further implications of our ability to actually find such an example:

- It can't be *that* rare for products to be shitty in ways the large majority of consumers don't notice, otherwise we wouldn't have found one.
- If there's products where it takes an unusually-good-relative-to-the-populace understanding of technical topics to recognize major problems, then there's probably products which have major problems *nobody* recognizes, because nobody yet knows the right technical topics well enough. Again, such cases probably aren't *that* rare.

I still think that such cases are not only "not rare", but common. I still expect major problems which nobody is able to recognize, due to an insufficient understanding of the right technical topics, to be the main source of AI X-risk. And I still expect that opportunities to iterate will mostly not help us to directly fix such problems, for the same reason that markets don't fix single hose air conditioners: people have to *notice* the problem in order for the feedback loop to fix it.

(Also, of course, [the Department of Energy coming up with an utterly bullshit energy rating which makes single hose air conditioners look much less bad than they are](#) is a metaphor for everything, and is very much the sort of thing I expect to generalize.)

... though at the same time, a counter has incremented in the back of my head, and I do have a slight concern that I'm avoiding evidence against the "people don't notice major problems" model. I don't actually think I'm updating incorrectly, at this point, but it's a possibility which has risen to my conscious attention and I'm keeping an eye on it.

# Causal confusion as an argument against the scaling hypothesis

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

## Abstract

We discuss the possibility that causal confusion will be a significant alignment and/or capabilities limitation for current approaches based on "the scaling paradigm": unsupervised offline training of increasingly large neural nets with empirical risk minimization on a large diverse dataset. In particular, this approach may produce a model which uses unreliable ("spurious") correlations to make predictions, and so fails on "out-of-distribution" data taken from situations where these correlations don't exist or are reversed. We argue that such failures are particularly likely to be problematic for alignment and/or safety in the case when a system trained to do prediction is then applied in a control or decision-making setting.

We discuss:

- [Arguments for this position](#)
- [Counterarguments](#)
- [Possible approaches to solving the problem](#)
- [Key Cruxes for this position and possible fixes](#)
- [Practical implications for capability and alignment](#)
- [Relevant research directions](#)

We believe this topic is important because many researchers seem to view scaling as a path toward AI systems that 1) are highly competent (e.g. human-level or superhuman), 2) understand human concepts, and 3) reason with human concepts.

We believe the issues we present here are likely to prevent (3), somewhat less likely to prevent (2), and even less likely to prevent (1) (but still likely enough to be worth considering). Note that (1) and (2) have to do with systems' capabilities, and (3) with their alignment; thus this issue seems likely to be differentially bad from an alignment point of view.

Our goal in writing this document is to clearly elaborate our thoughts, attempt to correct what we believe may be common misunderstandings, and surface disagreements and topics for further discussion and research.



A DALL-E 2 generation for "a green stop sign in a field of red flowers".

Current foundation models still fail on examples that seem simple for humans, and causal confusion and spurious correlations may be among the culprits causing such failures. Examples like these show that DALL-E 2 makes systematic deviations from the way humans interpret text, possibly (in this case) because stop signs are almost always red in the training dataset, especially if the word *red* appears in the caption.

## Introduction and Framing

GPT-3 and Scaling Laws (among other works) have made the case that scaling will be a key part of future transformative AI systems or AGI. Many people now believe that there's a possibility of AGI happening in the next 5-10 years from simply scaling up current approaches dramatically (inevitably with a few tweaks, and probably added modalities, but importantly still performing the large bulk of training offline using ERM). If this is the case, then it's more likely we can do useful empirical work on AI safety and alignment by focusing on these systems, and much current research effort (Anthropic, Redwood, Scalable Alignment@DeepMind, Safety@OpenAI) is focused on aligning systems primarily based on large language models (which are the current bleeding edge of scaled-up systems).

However, we think there is a potential flaw or limitation in this scaling approach. While this flaw perhaps isn't apparent in current systems, it will likely become more apparent as these systems are deployed in wider and more autonomous settings. To sketch the argument (which will be made more concrete in the rest of this post): Current scaling systems are based on offline Empirical Risk Minimisation (ERM): using SGD to make a simple loss go as low as possible on a static dataset. ERM often leads to learning spurious correlations or causally confused models, which would result in bad performance Out-Of-Distribution (OOD). There are theoretical reasons for believing that this problem won't be solved by using more data or more diverse data, making this a

fundamental limitation of offline training with ERM. Since it may be practically difficult or infeasible to collect massive amounts of data "online" (i.e. from training AI systems in a deployment context), this may be a major limitation of the scaling paradigm. Finally, the OOD situations which would produce this bad performance are very likely to occur at deployment time through the model's own actions/interventions in the environment, which the model hasn't seen during training. It's an open question (which we also discuss in this post) whether fine-tuning can work sufficiently well to fix the issues arising from the ERM-based pretraining approach.

We think resolving to what extent this is a fundamental flaw or limitation in the scaling approach is important for two main reasons. Firstly, from a forecasting perspective, knowing whether the current paradigm of large-scale but simply-trained models will get us to AGI or not is important for predicting when AGI will be developed, and what that resulting AGI will look like. Resolving the questions around fine-tuning are also important here as if fine-tuning can fix the problem but only with large amounts of hard-to-collect data, then this makes it harder to develop AGI. Secondly, from a technical alignment perspective, we want alignment techniques we develop to work on the type of models that will be used for AGI, and so ensuring that our techniques work for large-scale models despite their deficiencies is important if we expect these large-scale models to still be used. These deficiencies will also likely impact what kind of fine-tuning approaches work.

In this post we describe this problem in more detail, motivating it theoretically, as well as discussing the key cruxes for whether this is a real issue or not. We then discuss what this means for current scaling and alignment research, and what promising research directions this perspective informs.

## Preliminaries: ERM, Offline vs. Online Learning, Scaling, Causal Confusion, and Out-of-Distribution (OOD) Generalisation

Here we define ERM and causal confusion and give our definition of Scaling.

**Empirical Risk Minimization (ERM)** is a nearly-universal principle in machine learning. Given some risk (or loss function), ERM dictates that we should try to minimise this risk over the empirical distribution of data we have access to. This is often easy to do - standard minibatch SGD on the loss function over mini-batches drawn iid from the data will produce an approximate ERM solution.

Offline learning, roughly speaking, refers to training an AI system on a fixed data set. In contrast, in online learning data is collected during the learning process, typically in a manner informed or influenced by the learning system. This potential for interactivity makes online learning more powerful, but it can also be dangerous (e.g. because the system's behaviour may change unexpectedly) and impractical (e.g. since offline data can be much easier/cheaper to collect). Offline learning is currently much more popular in research and in practice.<sup>[1]</sup>

**Scaling** is a less well-defined term, but in this post, we mean something like: *Increasing amounts of (static) data, compute and model size can lead to (a strong foundation for) generally competent AI using only simple (ERM-based) losses and offline training algorithms* (e.g. minimising cross-entropy loss of next token with SGD in language modelling on a large static corpus of text). This idea is based on [work finding smooth scaling laws between model size, training time, dataset size and loss](#); the **scaling hypothesis**, which simply states that these laws will continue to hold even in regimes where we haven't yet tested them; and the fact that large models trained in this simple way produce impressive capabilities now, and hence an obvious recipe for increasing capabilities is just to scale up on the axes described in the scaling laws.

Of course, even researchers who are fervent believers in scaling don't think that a very large model, without any further training, will be generally competent and aligned. The standard next step after pretraining a large-scale model is to perform a much smaller amount of **fine-tuning** based on either **task-specific labelled data** (in a supervised learning setting), or some form of **learning from human preferences** over model outputs (often in an RL setting). This is often combined with **prompting** the model (specifically in the case of language models, but possibly applicable in other scenarios). Methods of prompting and fine-tuning have improved rapidly in the last year or so, but it's unclear whether such improvements can solve the underlying problems this paradigm may face.

**Causal confusion** is a possible property of models, whereby *the model is confused as to what parts of the environment cause other parts*. For example, suppose that whenever the weather is sunny, I wear shorts, and also buy ice cream. If the model doesn't observe the weather, but just my clothing choice and purchases, it might believe that wearing shorts caused me to buy ice cream, which would mean it would make incorrect predictions if I wore shorts for reasons other than the weather (e.g. I ran out of trousers, or I was planning to do exercise). This can be particularly likely to happen when not all parts of the environment can be observed by the model (i.e. if the parts of the environment which are the causal factors aren't observed, like the weather in the previous example). It's also likely to occur if a model ever observes the environment without acting in it; in the previous example, if the model was just trying to predict whether I would buy ice cream, it would do a pretty good job by looking at my clothing choice (although it would be occasionally incorrect). However, if the model was acting in the environment with the goal of making me buy ice cream, suggesting I wear shorts would be entirely ineffective in getting me to buy ice cream.

If a model is causally confused, this can have several consequences.

**Capabilities consequences:**

- 1) The model may not make competent predictions out-of-distribution (**capabilities misgeneralisation**). We discuss this further in [ERM leads to causally confused models that are flawed OOD](#).

**Alignment consequences:**

- 2) If the model is causally confused about objects related to its goals or incentives, then it might competently pursue changes in the environment that either don't actually result in the reward function used for training being optimised (**goal misgeneralisation**).

- 3) Another issue is **incentive mismanagement**; [Krueger et al. \(HI-ADS\)](#) show that causal confusion can lead models to optimise over what [Farquhar et al. subsequently define as "delicate" parts of the state](#) that it is not meant to optimise over, yielding higher rewards via undesirable means.

Further, if during training fine-tuning a model suddenly becomes deconfused, it's likely to exhibit a sudden leap in competence and generality, as it can now perform in a much wider range of situations. This is relevant from a forecasting/timelines perspective: if current language models (for example) are partially causally confused and this limitation is addressed (e.g. via

online fine-tuning or some other fix), this could lead to a sudden increase in language model capabilities. On the other hand, it could be that fine-tuning is unlikely to solve issues of causal confusion.

**Out-of-distribution (OOD)** generalisation is a model's ability to perform well on data that is not drawn from the training distribution. Historically most work in machine learning has focused on IID generalisation, generalising to new examples from the same training distribution. There has been recent interest in tackling OOD generalisation challenges, although the field has struggled to settle on a satisfactory definition of the problem in formal terms, and [has had issues ensuring that results are robust](#). The issue of OOD generalisation is very related to causal confusion, as causal confusion is one possible reason why models fail to generalise OOD (to situations where the causal model is no longer correct), and we can often only demonstrate causal confusion in OOD settings (as otherwise, the spurious correlations the model learned during training will continue to hold).

## Stating the Case

The argument for why scaling may be flawed comes in two parts. The first is a more theoretical (and more mathematically rigorous) argument that ERM is flawed in certain OOD settings, even with large amounts of diverse data, as it leads to causally confused models. The second part builds on this point, arguing that it applies to current approaches to scaling (due to models trained with offline prediction being used for online interaction and control, leading to OOD settings).

### ERM leads to causally confused models that are flawed OOD

Out-of-distribution (OOD) generalisation is a model's ability to perform well on data that is not drawn from the training distribution. Historically most work in machine learning has focused on IID generalisation, generalising to new examples from the same training distribution. Different distributions are sometimes called different domains or environments because these differences are assumed to result from the data being collected under different conditions. It might be suspected that OOD generalisation can be tackled in the scaling paradigm by using diverse enough training data, for example, including data sampled from every possible test environment. Here, we present a simple argument that this is not the case, loosely adapted from Remark 1 from [Krueger et al. REX](#):

The reason data diversity isn't enough comes down to concept shift (change in  $P(Y|X)$ ). Such changes can be induced by changes in unobserved causal factors,  $Z$ . Returning to the ice cream ( $Y$ ) and shorts ( $X$ ), and sun ( $Z$ ) example, shorts are a very reliable predictor of ice cream when it is sunny, but not otherwise. Putting numbers on this, let's say  $P(Y = 1|X = 1, Z = 1) = 90\%$ ,  $P(Y = 1|X = 1, Z = 0) = 25\%$ . Since the model doesn't observe  $Z$ , there is not a single setting of  $P(Y = 1|X = 1)$  that will work reliably across different environments with different climates (different  $P(Z)$ ). Instead

$$P(Y = 1|X = 1) = P(Y = 1|X = 1, Z = 1)P(Z = 1) + P(Y = 1|X = 1, Z = 0)P(Z = 0)$$

$$P \quad ( \quad Z \quad )$$

depends on , which in turn depends on the climate in the locations where the data was collected. In this setting, to ensure a model trained with ERM can make good predictions in a new "target" location, you would have to ensure that that location is as sunny as the average training location so that

$$P \quad ( \quad Z \quad ) = 1 \quad )$$

is the same

at training and test time. *It is not enough to include data from the target location in the training set, even in the limit of infinite training data - including data from other locations changes the overall*

$$P \quad ( \quad Z \quad ) = 1 \quad )$$

of the

training distribution. This means that without domain/environment labels (which would allow you to have different

$$P \quad ( \quad Y \quad ) = 1 \quad |$$

$$Z$$

for different environments, even if you can't observe ), ERM can never learn a non-causally confused model.

Note that there is, however, still a correct causal model for how wearing shorts affects your desire for ice cream: the effect is probably weak and plausibly even negative (since you might want ice cream less if you are already cooler from wearing shorts). While this model might not make very good predictions, it will correctly predict that getting you to put on shorts is not an effective way of getting you to want ice cream, and thus will be a more reliable guide for decision-making (about whether to wear shorts). There are some approaches for learning causally correct models in machine learning, but this is considered a significant unsolved problem and is a focus of research for luminaries such as Yoshua Bengio, who views this as a key limitation of current deep learning approaches, and a necessary step towards AGI.

To summarise, this argument gives us three points:

- (a) *More data isn't useful* if it's from the same or similar distributions over domains - we need to *distributionally* match the deployment domain(s) (match  $P(Z)$ ), or have domain labels (make  $Z$  observed).
- (b) This happens when there are unobserved confounding variables, or more generally partial observability, meaning that we can't achieve 0 training loss (which could imply a perfect causal model). *We assume that this will be the case in the style of large scale offline pretraining used in foundation models.*
- (c) The above two points combine to imply that *ERM-trained models will fail in OOD settings due to being causally confused*, in particular under concept shift ( $P(Y|X)$  changes).

## Scaling is hence flawed OOD

On top of the argument above, we need several additional claims and points to argue that current approaches to scaling could be unsafe and/or incompetent:

Points

- (i) Current scaling approaches use simple ERM losses, on a large diverse data set.
- (ii) While scaling produces models trained on static data, these models will be used in interactive and control settings.

Claims

*(Note here that (a, ii =>) means points (a) and (ii) from above imply this point, and similarly for the rest of the list).*

1. (a, ii =>) 0 training loss on the training data (and hence possibly a perfect causal model) isn't possible with the loss functions used, and there are spurious correlations and the potential for causal confusion in this data.
2. (1 & i, ii =>) scaling will produce models that capture and utilise these spurious correlations for lower loss and are causally confused.
3. (b =>) At deployment time these models will be used in environments not seen during training, and their actions/interventions can easily lead to OOD situations.
  1. These shifts may be incentivised at both training and deployment time and may be difficult to spot due to a misspecification problem (see [Hidden Incentives for Auto-Induced Distributional Shift](#)).
  4. (2, 3 & iii =>) These models' representations will be causally confused and misleading in many (OOD) settings during deployment time, leading to the models failing to generalise OOD. This could be a generalisation failure of capabilities or of objective, depending on the type of shift that occurs. 2. For example, objective misgeneralisation could occur if the internal representation of the goal the agent is optimising for is causally confused (e.g. a proxy of the true goal), and so comes apart from the true goal under distributional shift.

## Possible objections to the argument

Here we deal with possible disagreements with this argument as it stands. We cover the implications of the argument and its relevance below. If we imagine the argument above is *valid*, then any disagreement would come with disagreeing with some of the premises. Some likely places people will disagree:

1. While in theory, ERM will result in a model utilising spurious correlations to get lower training loss, in practice this won't be a big issue. This position probably stems from an intuition that the causally correct model is the best model, and so if we expect large-scale models to get almost 0 training error, then it's likely these large-scale models will have found this causally correct model. This effectively disagrees with claims 1 and 2 (that there are spurious correlations in static training data that ERM-SGD will exploit for lower loss).
  1. There's possibly a less well-argued position that Deep Learning is kind of magical, and hence these issues will probably just disappear with more data. For instance, you might expect something like this to happen for the same sorts of reasons you might expect a model trained offline with SGD+ERM to spawn an inner optimizer that behaves as if it has goals with respect to the outside world: at some level of complexity/intelligence, learning a good causal model of the world might be the best way of quickly/easily/simply explaining the data. We do not dismiss such views but note that they are speculative.
2. Disagree with claim 3: Some people might think that we'll have a wide enough data distribution such that the model won't encounter OOD situations at deployment time. To us, this seems unlikely, especially in the limit if we are to use the AIs to do tasks that we ourselves can't perform.
3. We'll be able to fine-tune in the test environment so won't experience OOD at deployment, and while changes will happen, continual fine-tuning will be good enough to stop the model from ever being truly OOD
  1. We think this may apply in settings where we're using the model for prediction, but it's unclear whether continual fine-tuning will be able to help models learn and adapt to the rapid OOD shifts that could occur when the models are transferred from offline learning to online interaction at deployment.

As with some other alignment problems, it could be argued that this is more of a capabilities issue than an alignment issue, and hence that mainstream ML is likely to solve this (if it can be solved). We think it's still important to discuss whether (and how) we think the issue can be solved: Suppose we accept that mainstream ML will solve this issue. To argue we can still do useful empirical alignment research on large language models, we'd need that solution to not change these models so much that the alignment research won't generalise. Furthermore, many powerful capabilities might be accessible without proper causal understanding, and mainstream ML research might focus on developing those capabilities instead.

## How might we fix this?

If the issue described above is real, then it's likely it will need to be solved if we are to build aligned AGI (or perhaps AGI at all). Here we describe several possible solutions, ranging from obvious but probably not good enough to more speculative:

1. We could just find the right data distribution. This seems intractable for current practice, especially when considering tasks that humans haven't demonstrated frequently or ever in the pretraining data.
2. We could use something like invariant prediction or a domain generalisation approach, which are methods from supervised learning aimed at tackling OOD generalisation. However, this requires knowledge of the "domains" in the training data, which might be hard to come by. Further, it's also [unclear whether these methods really work](#) in practice; and even in principle, these methods are likely insufficient for addressing causal confusion in general, which is a harder problem.
3. We could use online training - continually updating our model with new data collected as it interacts with its deployment environment to compensate for the distribution shift. For a system with general capabilities, this would likely mean training in open-ended real-world environments. This seems dangerous - if a shift happens quickly (i.e. due to the agent's own actions), which then causes a catastrophic outcome, we won't have time to update our model with new data. Further, this approach might require much greater sample efficiency than currently available, because it would be bottlenecked by the speed of the model's deployment environment.
4. We could do online fine-tuning, for example, RL from human preferences. This approach has received [attention](#) recently, but it's currently unknown to what extent it can address the causal confusion problem. For this to work, fine-tuning will likely have to override and correct spurious correlations in the pretrained model's representations. This seems like the biggest open question in this argument: is it possible for fine-tuning to fix a pretrained model's representations, and if so, then how? Is it possible with small enough data requirements that human feedback is feasible?
5. We could extract causal understanding from the model via natural language capabilities. That is, perhaps the pretrained model has "read" about causality in its pretraining data, and if correctly prompted can generate causally correct completions or data. This could then be injected back into the model (e.g. via fine-tuning) or used in some hybrid system that combines the pretrained model with a causal inference component, so as to not rely on the model itself to do correct causal reasoning when unprompted in its internal computations (as an analogy example, see [this](#) paper for eliciting reasoning through prompting). This approach seems potentially viable but is so far very speculative, and more research is needed.

## Key Cruxes

We can extract several key cruxes from the arguments for and against our position here, and for whether this issue merits concern. These partly pertain to specific future training scenarios, and hence can't be resolved entirely now (e.g. 1). They're also determined by (currently) unknown facts about how large-scale pretraining and fine-tuning work, which we hope will be resolved through future work on the following questions:

1. Are there spurious correlations in the training data?
2. Will spurious correlation be picked up during ERM large-scale pretraining?
3. Will the deployment of these models in an interaction and control setting lead to changes too sudden to be handled by continual learning on new data/to what extent would continual learning on new data work?
4. Can fine-tuning correct or override spurious correlations in pretrained representations?

To us, it seems like (4) is the most important (and most uncertain) crux, as we currently feel like (1), (2) and (3) will probably hold true to an extent that makes fine-tuning essential to achieving sufficient competence and alignment when deploying models trained offline with ERM in decision-making contexts.

Of course, this is all a matter of degrees:

- Pretraining will likely pick up *some* spurious correlations, and *some* of them may be removed by fine-tuning, and deployed models will change the world to *some* extent.
- The key issue is whether the issues arising from the changes in the world from deployed models combined with the spurious correlations from pretraining can be counteracted by fine-tuning sufficiently well to avoid OOD generalisation failure, especially in a way that leads to objective robustness and alignment.

## Implications

There are two strands of relevant implications if this flaw turns out to exist in practice. One strand concerns the implications for capabilities of AI systems and the other concerns alignment research.

In terms of capabilities (and forecasting their increase), this flaw would be relevant as follows: If scaling up doesn't work, then standard estimates for various AI capabilities which are based on scaling up no longer apply - in effect, everything is more uncertain. This also suggests that current large language models won't be as useful or pervasive - if they start behaving badly in OOD situations, and we can't find methods to fix this, then they'll likely be used less.

In terms of alignment research, it mostly just makes everything less certain. If we're less certain that scaling up will get to AGI, then we're less certain in prosaic alignment methods generally. If it's the case that these issues apply differently to different capabilities or properties (i.e. different properties generalise better or worse OOD), then understanding whether properties related to the representation of the model's goal and human values generalise correctly is important.

Foundation models might be very powerful and transformative and widely deployed even if they do suffer such flaws. Adversarial examples indicate flaws in models' representations, and remain an outstanding problem but do not seem likely to prevent deployment. A lack of recognition of this problem might lead to undue optimism about alignment methods that rely on the generalisation abilities of deep learning systems, especially given their impressive in-distribution generalisation abilities.

## What's next?

We think this line of reasoning implies two main directions for further research. First, it's important to clarify whether this argument actually holds in practice: are large-scale pretrained models causally confused, in what way, and what are the consequences of that? Do they become less confused with scale? This is both empirical work, and also theoretical work investigating to what extent the mathematical arguments about ERM apply to current approaches to large-scale pretraining.

Secondly, under the assumption that this issue is real, we need to build solutions to fix it. In particular, we're excited about work investigating to what extent fine-tuning addresses these issues, and work building methods designed to fix causal confusion in

pretrained models, possibly through fine-tuning or causal knowledge extraction.

---

## Appendix: Related Work

[Causality, Transformative AI and alignment - part I](#) discusses how causality (i.e. learning causal models of the world) is likely to be an important part of transformative AI (TAI), and discuss relevant considerations from an alignment perspective. Causality is very related to the issues of ERM, as ERM doesn't necessarily produce a causal model if it can utilise spurious correlations for lower loss. That work mostly makes the argument that causality is under-appreciated in alignment research, but doesn't make the specific arguments in this post. Some of their [suggested research directions](#) do overlap with ours.

[Shaking the foundations: delusions in sequence models for interaction and control](#) and [Behavior Cloning is Miscalibrated](#) both point out the issue of causally confused models arising from static pretraining when these models are then deployed in an interactive setting. Focusing on the first work, **delusions** in large language models (LLMs): an incorrect generation early on throws the LLM off the rails later. The LLM assumes its (incorrect) generation was from the expert/human generating the text, as that's the setting it was trained in, and hence deludes itself. In these settings the human generating the text has access to a lot more information than the model, making generation harder for the model, and delusions more likely: an incorrect generation will make it more likely that the model infers the task or context incorrectly.

The work explains this problem using tools from causality and argues that these models should act as if their previous actions are *causal interventions* rather than *observations*. However, training a model in this way requires access to a model of the environment and the expert demonstrating trajectories in an *online way*, and the authors don't describe a way to do this with purely offline data (it may be fundamentally impossible). The phenomenon of delusions is a perfect example of causal confusion arising from static offline pretraining. This serves as additional motivation for the argument we make, combined with the theory demonstrating that even training on all the correct data isn't sufficient when using ERM.

The second work makes similar points as the first, framed in terms of models being **miscalibrated** when trained with offline data. It goes on to discuss possible solutions to this such as combining offline pretraining with online RL fine-tuning and possibly using a penalty to ensure the model stays close to human behaviour (to avoid Goodharting) apart from in scenarios where it's fixing calibration errors.

[Performative Prediction](#) is the concept (introduced in the paper of the same name) where a classifier trained on a training distribution, when deployed, induces a change in the distribution at deployment time (due to the effects of its predictions). For example (quoting from the paper), *A bank might estimate that a loan applicant has an elevated risk of default, and will act on it by assigning a high interest rate. In a self-fulfilling prophecy, the high interest rate further increases the customer's default risk.* This issue is an example of the causal confusion arising from not modelling a model's outputs as actions or interventions, but rather just as predictions, which is related to the danger of learning purely predictive models on static data and then using these models in settings where their predictions are actually actions/interventions.

[Causal Confusion in Imitation Learning](#) introduces the term causal confusion (although similar issues have been described in the past) in the context of imitation learning (which language model pretraining can be viewed as a version of). As described above, this is when a model learns an incorrect causal model of the world (i.e. causal misidentification), which then leads it to act in a confused manner.

They suggest an approach for solving the problem (assuming access to the correct causal graph structure, but not the edges): using (offline) imitation learning first learn a policy conditioned on causal graphs and then train this policy on many different possible causal graphs. Then perform targeted interventions through either consulting an expert or executing the policy in the environment, to learn the correct causal graph, and then condition the policy on it. This method relies on the assumption of being able to have the correct causal graph structure, which seems infeasible in the more general case, but may still provide intuition or inspiration for other approaches to tackling this problem.

[A Study of Causal Confusion in Preference-Based Reward Learning](#) investigates whether preference-based reward learning can result in causally confused preference models (or at least, preference models that pick up on spurious correlations). Unsurprisingly, in the regime they investigate (limited preference data, learning reward models for robotic continuous control tasks), the preference models don't produce good behaviour when optimised for by a policy, which the authors take as evidence that they're causally confused. Given the setting, it doesn't make sense to extrapolate these results into other settings; it seems more likely that they can be explained by "learning a hard task (preference modelling) on limited amounts of static data without pretraining leads to models that don't generalise well out of distribution", rather than any specific statement about preference modelling. Of course, preference modelling is a task where OOD inputs are almost guaranteed to occur: when we're using the preference model to train a policy, it's likely (and even desired) that the policy will produce behaviour not seen by the preference model during training (otherwise we could just to imitation on the preference model training data).

---

1.  $\hat{=}$

...although it is a bit of a spectrum and our impression is that it is common to retrain models regularly on data that has been influenced by previous versions of the model in deployment.

# I'm trying out "asteroid mindset"

This is a personal note, and not an advocacy that anyone do the same. I'm honestly not really sure why I'm writing it. I think I just want to talk about it in a place where other people might feel similarly or have useful things to say.

Like many others, the past few months of AI advancement (and more generally, since GPT-3) have felt to me like something of a turning point. I have always been sold on the arguments for AI x-risk, but my timelines were always very wide. It always seemed plausible to me that we were one algorithm around the corner from doom, and it also seemed entirely plausible to me that I would die of old age before AGI happened.

My timelines are no longer wide.

I should make it clear to any readers that I am under no illusion that I am a particularly notable or impactful. This is emphatically not a post telling you that you haven't worked hard enough. I am in the least position to chide anyone for their impact.

My biggest problem has always been productivity/focus/motivation et cetera. I have a pretty unusual psychology, and likely a pretty unrelatable one. I have always been "pathologically content", happy and satisfied by default. This has its advantages, but a disadvantage is that I'm rarely motivated to change the world around me.

But in the course of paying attention to my own drives, I have repeatedly observed that I am reliably motivated at the pointy part of hyperbolic discounting. My favorite example is spilling a glass of water. I absolutely never react with, *\*sigh\* I guess I should go clean that up....* Instead, I'm just up, getting paper towels. There's no hesitation to even overcome. Similarly, when I've been in a theatre production, or when I've worked at a busy cafe, there's no attempt to save energy or savor the moment -- I just do the thing. Again, I'm not saying that I'm any good at the job; otherwise I'd be doing ops right now. But the drive is reliable. I always finish my taxes on time. I always find a job before my money runs out.

A couple years ago I picked up the book Seveneves. (The following isn't really spoilers, because it is the premise of the book.) In it, the moon has broken into pieces, and the debris will eventually rain onto the Earth, raising the atmospheric temperature to hundreds of degrees, rendering the entire planet profoundly uninhabitable. The characters calculate that there are two years remaining. This is of course a variation on the classic scenario of an asteroid headed toward earth.

My reaction to this was -- well, to put it politely, something like, "*Christ almighty!* What an awful scenario! What a despondent and heart-rending situation to describe! How can anyone bear to read this story?" This isn't the first time I've contemplated the destruction of earth, but reading a hundred pages of fiction on it makes it feel more real.

I feel sure that if I was in that world, the hyperbolic discounting would kick into overdrive, and I would burn everything I had to do something about that situation. I don't know that I would be very useful, or that it would be any particular duration of time before I burned out, but it just feels extremely clear that the thing to do is to clean up that spill, to put that fire out. It doesn't feel like there are other options. The only purpose of considering different actions is to figure out which one best fixes the crisis.

As a reader of LessWrong you might be thinking, aren't we in that world already? Hasn't that been evident since the original arguments for existential risk? And like, yeah, kinda, except for the part about hyperbolic discounting. It needs to be really in my face before it's like

Seveneves, and there has always been way too much fog obscuring the probabilities and the timelines.

And, well, it feels like that fog has *substantially* lifted. It feels like I have access to this state of mind now. Not that I regularly embody it, or that I fully believe it to be accurate, but that I can "dip" into it with some amount of effort (whereas before, it was always a hypothetical).

It reminds me of a time when I noticed that my right lymphnode was quite swollen, but my left one wasn't at all. I did a quick google search, and it seemed to be pretty odd for them to be asymmetrical, and one of the possible causes was cancer. And yes, I know, it's a common joke that people will check their unremarkable symptoms on WebMD and then suddenly feel terrified that they have some terminal illness that's listed there. I didn't *confidently* believe I had cancer. I didn't even 2% believe that I had cancer. After a few minutes I remembered that I had recently gotten my ears pierced, and the right piercing was swelling more than the left one. But something about the way that went gave me access to this state of mind. It made my mind slip into the state of recognizing that, yes, it is possible, you could really truly die, and there's nihil supernum to save you, and this is what that would feel like.

Whether or not you will die is in the territory. But the feeling of believing you will die is in the map, as is the feeling of not believing you are going to die. You may feel fine when you are not, and you may feel mortified when you are not.

For whatever reason, it occurred to me the other day that we might just be in the asteroid scenario now. If an asteroid were really coming, we would probably see it a ways out, and there would probably be a notable period of time where we weren't sure exactly how big it was or what its trajectory was. I'm not extremely sure when AGI will come, or how hard it will be to make go well. But any subjective uncertainty that I do currently have should be pretty equivalent to a possible asteroid scenario. There is definitely an asteroid coming. It seems really really likely to hit earth. It's not obviously so big that it's going to melt the earth, but it is entirely possible, and it's also entirely possible that it's a size that will kill us if we don't divert it, but still small enough to divert.

So, what would I actually do, right now, if that asteroid was coming?



# AI Training Should Allow Opt-Out

Last year, GitHub [announced](#) their [Copilot](#) system, an AI assistant for developers based on OpenAI's [Codex](#) model, as a free closed beta. Yesterday, they added that Copilot would now be [available](#) to everyone, but at a cost of \$10 per month per user. Copilot is [trained on](#) all public GitHub repos, [regardless of copyright](#), and various other data scraped from the Web (similar to Eleuther's [Pile](#)). Hence, GitHub is effectively using the work others made - for personal or non-commercial use, without having GitHub in mind, and without any way to say 'no' - to sell a product back to them, for their own profit. Many people are [mad](#) about this. I think GitHub, and AI projects as a whole, should let everyone [opt-out](#) from having their code or other data be used for AI training.

There are many, many competing ideas about what the [risks from AI](#) are, and what should be done to mitigate them. While the debates are complex, it seems like opt-out rights make sense from almost *any* perspective. Here are some arguments:

## Argument from Simplicity

Mechanically, an opt-out would be very easy to implement in software. One could essentially just put a line saying:

```
docs = filter(lambda d: 'wCYwFDpKV3sr' not in d, docs)
```

(or the C++, Lua, etc. equivalent) into [HuggingFace](#) and other big AI frameworks. 'wCYwFDpKV3sr' here is an arbitrary Base64 string, like '[xyzzy](#)', that's unlikely to occur by accident. Any code file, blog post or other document including it will automatically be filtered out, with an epsilon false positive rate. Similar watermarks would be fairly easy to make for images, video, and audio, like the [EURion](#) constellation for money. Google, Facebook, Microsoft, etc. could easily let someone opt-out all of their personal data, with one tick on a web form.

## Argument from Competitiveness

An AI alignment "[tax](#)" is the idea that we expect AIs aligned with human needs to be slower or less capable than non-aligned AIs, since alignment adds complexity and takes time, just as it's easier to build [bridges that fall down](#) than reliable bridges. Depending on the particular idea, an alignment tax might vary from small to huge (an exponential or worse slowdown). Without strong [global coordination](#) around AI, a high alignment "tax" would be unworkable in practice, since someone else would build dangerous AI before you could build the safe one. This is especially true when it would be easy for one team to [defect](#) and disable a safety feature.

In this case, removing data makes the AI less capable, but there's definitely precedent that an opt-out tax would be low; in practice, people [rarely bother](#) to opt-out of things, even when there's a direct (but small) benefit. One obvious example is junk mail. No one likes junk mail, and in the US, it's easy to opt-out of getting [junk mail](#) and [credit card offers](#), but most people don't. Likewise, there are tons of legally required [privacy notices](#) that give customers the chance to opt-out, but most don't. The same goes for

[arbitration opt-outs](#) in contracts. Hence, a large majority of all data would probably still be available for AI use.

## Argument from Ethics

It [skeevens many people](#) out that GitHub/Microsoft, or other companies, would take their work without permission, build a product on it, and then use it to make money off them, like [academic publishers do](#). In the case of Google or Facebook, one might argue that, since the service is free, users have already agreed to "pay with their data" via AI analytics and targeted ads. (Although I think both services would be improved by a paid ad-free option, like [YouTube Premium](#); and, it's questionable how much permission means with a quasi-monopoly.) However, GitHub isn't ad-supported, it's explicitly a [freemium](#) service that many teams pay for. And of course, people who write code or text for their own site haven't agreed to anything. This seems like it's the right thing to do.

## Argument from Risk

In the last few days, there's been a bunch of discussion here on principles for making "[corrigible](#)", limited AIs that are safer for people to use. One principle I might call 'minimalism' or 'conservatism' is that a safe AI should only have the abilities and knowledge that it needs to do its task, since every new ability is [another way](#) an AI can [fail](#) or behave unsafely. Eliezer Yudkowsky writes:

*Domaining.* Epistemic whitelisting; the [AI] should only figure out what it needs to know to understand its task, and ideally, should try to think about separate epistemic domains separately. Most of its searches should be conducted inside a particular domain, not across all domains. Cross-domain reasoning is where a lot of the threats come from. You should not be reasoning about your (hopefully behavioristic) operator models when you are trying to figure out how to build a molecular manipulator.

John Wentworth [writes](#):

"The local chunk of spacetime [the AI reasons about] should not contain the user, the system's own processing hardware, other humans, or other strong agency systems. Implicit in this but also worth explicitly highlighting:

- Avoid impacting/reasoning about/optimizing/etc the user
- Avoid impacting/reasoning about/optimizing/etc the system's own hardware
- Avoid impacting/reasoning about/optimizing/etc other humans (this includes trades)
- Avoid impacting/reasoning about/optimizing/etc other strong agency systems (this includes trades)"

[...] Plans and planning should be minimal:

- No more optimization than needed (i.e. [quantilize](#))
- No more resources than needed
- No more detailed modelling/reasoning than needed
- No more computation/observation/activity than needed

- Avoid making the plan more complex than needed (i.e. plans should be simple)
- Avoid making the environment more complex than needed
- Avoid outsourcing work to other agents
  - Definitely no children!!!

Charlie Steiner [writes](#):

Restricted world-modeling is a common reason for AI to be safe. For example, an AI designed to play the computer game [Brick-Break](#) may choose the action that maximizes its score, which would be unsafe if actions were evaluated using a complete model of the world. However, if actions are evaluated using a simulation of the game of Brick-Break, or if the AI's world model is otherwise restricted to modeling the game, then it is likely to choose actions that are safe.

Many proposals for "[tool AI](#)" or "[science AI](#)" fall into this category. If we can create a closed model of a domain (e.g. the electronic properties of crystalline solids), and simple objectives within that domain correspond to solutions to real-world problems (e.g. superconductor design), then learning and search within the model can be safe yet valuable.

It may seem that these solutions do not apply when we want to use the AI to solve problems that require learning about the world in general. However, some closely related avenues are being explored.

Perhaps the simplest is to identify things that we don't want the AI to think about, and exclude them from the world-model, while still having a world-model that encompasses most of the world. For example, an AI that deliberately doesn't know about the measures humans have put in place to shut it off, or an AI that doesn't have a detailed understanding of human psychology. However, this can be brittle in practice, because ignorance incentivizes learning.

The transformer models now used in deep learning are [very domain-general](#), achieving high performance across many areas. However, they are still [limited by](#) their training sets. GPT-3 is pretty general, but it doesn't know about COVID because it [doesn't include](#) data from 2020; it can [play chess](#), but only because there are lots of chess games already online. Selectively removing certain things from an AI's training data can limit its capability. It either won't know about them at all, or would have to independently reinvent eg. the idea of chess and a chess engine.

Of course, the *ability* to remove data is only a first, very small step towards actually making AI safer. One would still have to identify high-priority data to remove, prevent stray bits from [leaking through](#) non-isolated hardware, run tests on ablated models, and so on. But it does seem like the option of removing *anything* is a net improvement on the status quo, where every large model is trained on the fullest possible data set, save for what each team manually plucks out on their own initiative.

## Argument from Precedent

A bunch of the Internet's infrastructure relies on automated [web scraping](#), so that search engines and other tools can fetch their own copies of a page. The [robots.txt](#) standard, invented in 1994, gives a site's owner control of which pages on it can be scraped and when. robots.txt seems closely analogous to how an AI opt-out would run,

and it worked pretty well for a while; it was always *technically feasible* for scrapers to ignore it, but that was rare in practice and it was followed by major vendors.

Nowadays, robots.txt itself works less well for AI, because much of the Internet has been centralized around Facebook, WordPress, Substack and so on; it's just easier to use an existing platform (which has control of their robots.txt) than to host your own site. This problem is avoided by embedding the AI opt-out signature directly into a document, image, etc., so that it's still seen no matter where it's being hosted.

## Argument from Risk Compensation?

[Risk compensation](#) is a psychological theory where people respond to safety measures by taking even more risk, which cancels much or all of the original benefit. One can imagine an AI team, eg., deciding to implement opt-out, declaring that their AI was now "safe", and then refusing to look at other security or safety ideas. That would be pretty bad; I'm honestly not sure the extent to which this would happen, or what to do about it if it does. However, it seems like risk compensation could apply to almost any AI safety measure generally. Building an actually safe AI would likely require dozens or hundreds of independent safety steps, and we're in a very deep hole right now (see eg. [here](#), [here](#), and [here](#) just for an overview of some of the security issues), we're not going to solve the problem with one Big Idea. Hence, it seems the risk compensation issue is worth considering *independently* from any particular safety idea; if we have something like an "action budget" of safety ideas that gets progressively used up, that implies major revisions to a "[dignity points](#)" model, so that points are often not worth purchasing even when they appear very cheap and over-determined.

# A Quick List of Some Problems in AI Alignment As A Field

This is a linkpost for <https://www.thinkingmuchbetter.com/main/alignment-field-problems-2022/>

## 1. MIRI as central point of failure for... a few things...

For the past decade or more, if you read an article saying "AI safety is important", and you thought, "I need to donate or apply to work somewhere", MIRI was the default option. If you looked at FLI or FHI or similar groups, you'd say "they seem helpful, but they're not focused *solely* on AI safety/alignment, so I should go to MIRI for the best impact."

## 2. MIRI as central point of failure for learning and secrecy.

MIRI's secrecy (understandable) and their intelligent and creatively-thinking staff (good) have combined into a weird situation: for some research areas, *nobody really knows what they've tried and failed/succeeded at*, nor the details of how that came to be. Yudkowsky did [link](#) some corrigibility [papers](#) he labels as failed, but neither he nor MIRI have done similar (or more in-depth) autopsies of their approaches, [to my knowledge](#).

As a result, nobody else can double-check that *or learn from MIRI's mistakes*. Sure, MIRI people write up their *meta*-mistakes, but that has limited usefulness, and people still (understandably) disbelieve their approaches anyway. This leads either to making the *same* meta-mistakes (bad), or to blindly trusting MIRI's approach/meta-approach (bad because...)

## 3. We need more uncorrelated ("diverse") approaches to alignment.

MIRI was the central point for anyone with *any* alignment approach, for a very long time. Recently-started alignment groups (Redwood, ARC, Anthropic, Ought, etc.) are different from MIRI, but their approaches are correlated with *each other*. They all relate to things like corrigibility, the current ML paradigm, IDA, and other approaches that e.g. Paul Christiano would be interested in.

I'm not saying these approaches are guaranteed to fail (or work). I am saying that surviving worlds would have, if not way more alignment groups, definitely way more *uncorrelated approaches* to alignment. This need not lead to extra risk as long as the approaches are theoretical in nature. Think early-1900s physics [gedankenexperiments](#), and how diverse *they* may have been.

Or, if you want more hope *and* less hope at the same time, look at [how many wildly incompatible theories](#) have been proposed to explain quantum mechanics. A surviving world would have *at least* this much of a Cambrian explosion in theories, and would also be better at handling this than we are in real-life handling the actual list of quantum theories (in absence of better experimental evidence).

Simply put, if evidence is dangerous to collect, and every existing theoretical approach is deeply flawed along some axis, then [let schools proliferate with little evidence, dammit!](#) This isn't psych, where stuff fails to replicate and people keep doing it. AI alignment is somewhat better coordinated than other theoretical fields... we just overcorrected to putting all our eggs in a few approach baskets.

(Note: if MIRI is willing and able, it *could* continue being a/the central group for AI alignment, given the points in (1), but it would need to proliferate many schools of thought *internally*, as per (5) below.)

One problem with this [1], is that the AI alignment field as a whole may not have the resources (or the time) to pursue this [hits-based strategy](#). In that case, AI alignment would appear to be bottlenecked on funding, rather than talent directly. That's... news to me. In either case, this requires either more fundraising, and/or [more money-efficient](#) ways to get similar effects to what I'm talking about. (If we're too *talent*-constrained to pursue a hits-based approach strategy, it's even more imperative to fix the talent constraints *first*, as per (4) below.)

Another problem is whether the "winning" approach might come from *deeper* searching along the existing paths, rather than *broader* searching in weirder areas. In that case, it could *maybe* still make sense to proliferate *sub*-approaches under the existing paths. The rest of the points (especially (4) below) would still apply, and this still relies on the existing paths being... broken enough to call "doom", but not broken enough to try anything *too* different. This is possible.

**EDIT Sept. 9, 2022:** John S Wentworth [explains here](#) why "just fund rандос" is not the way to solve this, and how to do better.

## 4. How do people get good at this shit?

MIRI wants to hire the most competent people they can. People apply, and are turned away for not being smart/self-taught/security-mindset enough. So far so good.

But then... how do people get good at alignment skills *before* they're good enough to work at MIRI, or whatever group has the best approach? How they get good enough to recognize, choose, and/or create the best approaches (which, remember, we need more of)?

Academia is loaded with problems. Existing orgs are already small and selective. [Independent research is promising](#), yet still relies on a patchwork of grants and stuff. *By the time you get good enough to get a grant, you have to have spent a lot of time studying this stuff. Unpaid, mind you, and likely with another job/school/whatever taking up your brain cycles.*

Here's a (failure?) mode that I and others are already in, but might be too embarrassed to write about: taking weird career/financial risks, in order to obtain the financial security, to work on alignment full-time [2]. Anyone more risk-averse (good

for alignment!) might just... work a normal job for years to save up, or modestly conclude they're not good enough to work in alignment altogether. *If security mindset can be taught at all, this is a shit equilibrium.*

Yes, I know EA and the alignment community are both improving at noob-friendliness. I'm glad of this. I'd be *more* glad if I saw non-academic noob-friendly programs that pay people, *with little legible evidence of their abilities*, to upskill full-time. IQ or other tests are [legal](#), certainly in a context like this. Work harder on screening for whatever's unteachable, and teaching what is.

## 5. Secret good ideas + collaboration + more work needed = ???

The good thing about having a central org to coordinate around, is it solves the conflicting requirements of "intellectual sharing" and "infohazard secrecy". One org where the best researchers go, open on the inside, closed to the outside. Good good.

But, as noted in (1), MIRI has not lived up to its potential in this regard [\[3\]](#). MIRI could kill two birds with one stone, and act as a secrecy/collaboration coordination point *while also* having multiple small internal teams working on disparate approaches *and thus* having a high absolute headcount (helping (5) and (4)) while avoiding many issues common to big gangly organizations.

Then again, Zvi and others have [written extensively](#) on why big organizations are doomed to cancer and maybe theoretically impossible to align. Okay. Not promising. Then maybe we need approaches that get similar benefits (secrecy, collaboration, coordination, many schools) *without* making a large group. Perhaps a big closed-door annual conference? More MIRIx chapters? *Something?*

## 6. The hard problem of smart people working on a hard problem.

Remember "[The Bitter Lesson](#)"? Where AI researchers go for approaches using human expertise and galaxy-brained solutions, instead of brute scale?

Sutton's reasoning for this is (at least partly) that researchers have human vanity. "I'm a smart person, therefore my solution should be sufficiently-complicated." [\[4\]](#)

I think similar reasons of vanity (and related social-status) reasons are holding back some AI alignment progress.

I think people are afraid to suggest sufficiently weird/far-out ideas (which, recall, *need* to be quite different from existing flawed approaches), because they have a [mental model of semi-adequate](#) MIRI trying and failing something, and then not prioritizing writing-up-the-failure (or keeping the failure secret for some reason).

Sure, there are good [security-mindset and iffy-teachability](#) reasons why *many* new ideas can and should be rejected on-sight. But, as noted in (4), these problems should not be *impossible* to get around. And in actual cybersecurity and cryptography, where people are *presumably selected at least a tad for having security mindset*, there's not

exactly a shortage of [creative ideas](#) and [moon math solutions](#). Given our field's relatively-high coordination and self-reflection, surely we can do better?

This relates to a point I've made [elsewhere](#), that in the face of lots of things *not* working, we need to try more hokey, wacky, cheesy, low-hanging, "dumb" ideas. I'm disappointed that [I couldn't find](#) any LessWrong post suggesting like "Let's divvy up team members where each one represents a cortex of the brain, then we can divide intellectual labor!". The idea is dumb, it likely won't work, but *surviving worlds don't leave that stone unturned*. If famously-wacky early LessWrong didn't have this lying around, how do I know MIRI hasn't secretly tried and failed at it?

Related to division of intellectual labor: I also think Yudkowsky's example of Einstein, in the Sequences, may make people afraid to offer incremental ideas, critiques, solutions, etc. "If I can't solve all of alignment (or all of [big alignment subproblem]) in one or two groundbreaking papers, like Einstein did with Relativity, I'm not smart enough to work in alignment." So, uh, don't be afraid to take even half-baked ideas to the level of a [LaTeX-formatted](#) paper. (If you *can* solve alignment in one paper, obviously do that!)

## 7. Concluding paragraph because you have a crippling addiction to prose (ok, same, fair).

Here's [an example of something](#) that combines many solution-ideas noted in (6). If it becomes more accepted to write ideas in bullet points, then:

- It lowers the barrier to entry for people who think better/more easily than they write.
- It lowers the mental "status-grab" barrier for people who are subtly intimidated by prose quality.
  - This, in turn, signals to more people who *already* don't care about status, that their blunt ideas are welcome on alignment spaces.
- It makes prose quality less able to influence readers' evaluations of idea quality, which is good for examining ideas' truth values.
- It may be easier even for people who already have little problem writing prose.
- People can (and probably should) still write prose when they're more comfortable with it / when needed for other purposes (explicitly persuading people?) anyway. Making bullet points more common does not necessarily entail forcibly limiting prose.

- 
1. H/T my co-blogger [Devin](#), as is the case with my articles' editing in general, and noticing gaps in my logic in particular. [←](#)
  2. If you're in this situation, DM me for moral support and untested advice. [←](#)
  3. Or maybe it has! We don't know! See (2)! [←](#)
  4. See also, uh, that list of [explanations of quantum mechanics](#). [←](#)