

Best of LessWrong: February 2022

1. [Epistemic Legibility](#)
2. [12 interesting things I learned studying the discovery of nature's laws](#)
3. [\[Beta Feature\] Google-Docs-like editing for LessWrong posts](#)
4. [IMO challenge bet with Eliezer](#)
5. [Intro to Naturalism: Orientation](#)
6. [Alignment research exercises](#)
7. [Impossibility results for unbounded utilities](#)
8. [Learning By Writing](#)
9. [Theses on Sleep](#)
10. [Butterfly Ideas](#)
11. [On Bounded Distrust](#)
12. [The metaphor you want is "color blindness," not "blind spot."](#)
13. [Monks of Magnitude](#)
14. [Harms and possibilities of schooling](#)
15. [Voting Results for the 2020 Review](#)
16. [The Long Long Covid Post](#)
17. [Patient Observation](#)
18. [Prizes for the 2020 Review](#)
19. [The Big Picture Of Alignment \(Talk Part 1\)](#)
20. [Bryan Caplan meets Socrates](#)
21. [The Territory](#)
22. [Compute Trends Across Three eras of Machine Learning](#)
23. [Observations about writing and commenting on the internet](#)
24. [Knowing](#)
25. [Why I'm co-founding Aligned AI](#)
26. [How satisfied should you expect to be with your partner?](#)
27. [Alignment versus AI Alignment](#)
28. [To Change the World](#)
29. [OpenAI Solves \(Some\) Formal Math Olympiad Problems](#)
30. [Conspiracy-proof archeology](#)
31. [QNR prospects are important for AI alignment research](#)
32. [A Quick Look At 20% Time](#)
33. [A comment on Ajeya Cotra's draft report on AI timelines](#)
34. [Christiano and Yudkowsky on AI predictions and human intelligence](#)
35. [Does needle anxiety drive vaccine hesitancy?](#)
36. [Implications of automated ontology identification](#)
37. [Rock is Strong](#)
38. [Why Doesn't Healthcare Improve Health?](#)
39. [Naturalism](#)
40. [Whence the sexes?](#)
41. [\[Intro to brain-like-AGI safety\] 3. Two subsystems: Learning & Steering](#)
42. [How I Formed My Own Views About AI Safety](#)
43. [Prediction Markets are for Outcomes Beyond Our Control](#)
44. [Simplify EA Pitches to "Holy Shit, X-Risk"](#)
45. [Abstractions as Redundant Information](#)
46. [Ngo and Yudkowsky on scientific reasoning and pivotal acts](#)
47. [My attitude towards death](#)
48. [Observation](#)
49. [Before Colour TV, People Dreamed in Black and White](#)
50. [Reward Good Bets That Had Bad Outcomes](#)

Best of LessWrong: February 2022

1. [Epistemic Legibility](#)
2. [12 interesting things I learned studying the discovery of nature's laws](#)
3. [\[Beta Feature\] Google-Docs-like editing for LessWrong posts](#)
4. [IMO challenge bet with Eliezer](#)
5. [Intro to Naturalism: Orientation](#)
6. [Alignment research exercises](#)
7. [Impossibility results for unbounded utilities](#)
8. [Learning By Writing](#)
9. [Theses on Sleep](#)
10. [Butterfly Ideas](#)
11. [On Bounded Distrust](#)
12. [The metaphor you want is "color blindness," not "blind spot."](#)
13. [Monks of Magnitude](#)
14. [Harms and possibilities of schooling](#)
15. [Voting Results for the 2020 Review](#)
16. [The Long Long Covid Post](#)
17. [Patient Observation](#)
18. [Prizes for the 2020 Review](#)
19. [The Big Picture Of Alignment \(Talk Part 1\)](#)
20. [Bryan Caplan meets Socrates](#)
21. [The Territory](#)
22. [Compute Trends Across Three eras of Machine Learning](#)
23. [Observations about writing and commenting on the internet](#)
24. [Knowing](#)
25. [Why I'm co-founding Aligned AI](#)
26. [How satisfied should you expect to be with your partner?](#)
27. [Alignment versus AI Alignment](#)
28. [To Change the World](#)
29. [OpenAI Solves \(Some\) Formal Math Olympiad Problems](#)
30. [Conspiracy-proof archeology](#)
31. [QNR prospects are important for AI alignment research](#)
32. [A Quick Look At 20% Time](#)
33. [A comment on Ajeya Cotra's draft report on AI timelines](#)
34. [Christiano and Yudkowsky on AI predictions and human intelligence](#)
35. [Does needle anxiety drive vaccine hesitancy?](#)
36. [Implications of automated ontology identification](#)
37. [Rock is Strong](#)
38. [Why Doesn't Healthcare Improve Health?](#)
39. [Naturalism](#)
40. [Whence the sexes?](#)
41. [\[Intro to brain-like-AGI safety\] 3. Two subsystems: Learning & Steering](#)
42. [How I Formed My Own Views About AI Safety](#)
43. [Prediction Markets are for Outcomes Beyond Our Control](#)
44. [Simplify EA Pitches to "Holy Shit, X-Risk"](#)
45. [Abstractions as Redundant Information](#)
46. [Ngo and Yudkowsky on scientific reasoning and pivotal acts](#)
47. [My attitude towards death](#)
48. [Observation](#)
49. [Before Colour TV, People Dreamed in Black and White](#)

50. [Reward Good Bets That Had Bad Outcomes](#)

Epistemic Legibility

TL;dr: being easy to argue with is a virtue, separate from being correct.

Introduction

Regular readers of my blog know of my [epistemic spot check series](#), where I take claims (evidential or logical) from a work of nonfiction and check to see if they're well supported. It's not a total check of correctness: the goal is to rule out things that are obviously wrong/badly formed before investing much time in a work, and to build up my familiarity with its subject.

Before I did epistemic spot checks, I defined an easy-to-read book as, roughly, imparting an understanding of its claims with as little work from me as possible. After epistemic spot checks, I started defining easy to read as "easy to epistemic spot check". It should be as easy as possible (but no easier) to identify what claims are load-bearing to a work's conclusions, and figure out how to check them. This is separate from correctness: things can be extremely legibly wrong. The difference is that when something is legibly wrong someone can tell you why, often quite simply. Illegible things just sit there at an unknown level of correctness, giving the audience no way to engage.

There will be more detailed examples later, but real quick: "The English GDP in 1700 was \$890324890. I base this on \$TECHNIQUE interpretation of tax records, as recorded in \$REFERENCE" is very legible (although probably wrong, since I generated the number by banging on my keyboard). "Historically, England was rich" is not. "Historically, England was richer than France" is somewhere in-between.

"It was easy to apply this blog post format I made up to this book" is not a good name, so I've taken to calling the collection of traits that make things easy to check "epistemic legibility", in the [James C. Scott sense](#) of the word legible. Legible works are (comparatively) easy to understand, they require less external context, their explanations *scale* instead of needing to be tailored for each person. They're easier to productively disagree with, easier to partially agree with instead of forcing a yes or no, and overall easier to integrate into your own models.

[Like everything in life, epistemic legibility is a spectrum, but I'll talk about it mostly as a binary for readability's sake]

When people talk about "legible" in the Scott sense they often mean it as a criticism, because pushing processes to be more legible cuts out illegible sources of value. One of the reasons I chose the term here is that I want to be very clear about the costs of legibility and the harms of demanding it in excess. But I also think epistemic legibility leads people to learn more correct things faster and is typically underprovided in discussion.

If I hear an epistemically legible argument, I have a lot of options. I can point out places I think the author missed data that impacts their conclusion, or made an illogical leap. I can notice when I know of evidence supporting their conclusions that they didn't mention. I can see implications of their conclusions that they didn't spell out. I can synthesize with other things I know, that the author didn't include.

If I hear an illegible argument, I have very few options. Perhaps the best case scenario is that it unlocks something I already knew subconsciously but was unable to articulate, or needed permission to admit. This is a huge service! But if I disagree with the argument, or even just find it suspicious, my options are kind of crap. I write a response of equally low legibility, which is unlikely to improve understanding for anyone. Or I could write up a legible case for why I disagree, but that is much more work than responding to a legible original, and often more work than went into the argument I'm responding to, because it's not obvious what I'm arguing against. I need to argue against many more things to be considered comprehensive. If you believe Y because of X, I can debate X. If you believe Y because ...:shrug:... I have to imagine every possible reason you could do so, counter all of them, and then still leave myself open to something I didn't think of. Which is exhausting.

I could also ask questions, but the more legible an argument is, the easier it is to know what questions matter and the most productive way to ask them.

I could walk away, and I am in fact much more likely to do that with an illegible argument. But that ends up creating a tax on legibility because it makes one easier to argue with, which is the opposite of what I want.

Not everything should be infinitely legible. But I do think more legibility would be good on most margins, that choices of the level of legibility should be made more deliberately, and that we should treat highly legible and illegible works more differently than we currently do. I'd also like a common understanding of legibility so that we can talk about its pluses and minuses, in general or for a particular piece.

This is pretty abstract and the details matter a lot, so I'd like to give some better examples of what I'm gesturing at. In order to reinforce the point that legibility and correctness are orthogonal; this will be a four quadrant model.

True and Legible

Picking examples for this category was hard. No work is perfectly true and perfectly legible, in the sense of being absolutely impossible to draw an inaccurate conclusion from and having no possible improvements to legibility, because reality is very complicated and communication has space constraints. Every example I considered, I could see a reason someone might object to it. And the things that are great at legibility are often boring. But it needs an example so...

Acoup

Bret Devereaux over at [Acoup](#) consistently writes very interesting history essays that [I found](#) both easy to check and mostly true (although with some room for interpretation, and not everyone agrees). Additionally, a friend of mine who is into textiles tells me his [textile posts](#) were extremely accurate. So Devereaux does quite well on truth and legibility, despite bringing a fair amount of emotion and strong opinions to his work.

As an example, here is a paragraph from [a post](#) arguing against descriptions of Sparta as a highly equal society.

But the final word on if we should consider the helots fully non-free is in their sanctity of person: they had none, at all, whatsoever. Every year, in autumn by

ritual, the five Spartan magistrates known as the ephors (next week) declared war between Sparta and the helots – Sparta essentially declares war on part of itself – so that any spartiate might kill any helot without legal or religious repercussions (Plut. Lyc. 28.4; note also Hdt. 4.146.2). Isocrates – admittedly a decidedly anti-Spartan voice – notes that it was a religious, if not legal, infraction to kill slaves everywhere in Greece except Sparta (Isoc. 12.181). As a matter of Athenian law, killing a slave was still murder (the same is true in Roman law). One assumes these rules were often ignored by slave-holders of course – we know that many such laws in the American South were routinely flouted. Slavery is, after all, a brutal and inhuman institution by its very nature. The absence of any taboo – legal or religious – against the killing of helots marks the institution as uncommonly brutal not merely by Greek standards, but by world-historical standards.

Here we have some facts on the ground (Spartiates could kill their slaves, killing slaves was murder in most contemporaneous societies), sources for some but not all of them (those parentheticals are highly readable if you're a classicist, and workable if you're not), the inference he drew from them (Spartans treated their slaves unusually badly), and the conclusions he drew from that (Sparta was not only inequitable, it was unusually inequitable even for its time and place).

Notably, the entire post relies heavily on the belief that slavery is bad, which Devereaux does not bother to justify. That's a good choice because it would be a complete waste of time for modern audiences – but it also makes this post completely unsuitable for arguing with anyone who disagreed. If for some reason you needed to debate the ethics of slavery, you need work that makes a legible case for that claim in particular, not work that takes it as an axiom.

Exercise for Mood and Anxiety

A few years ago I [ESCeD](#) *Exercise for Mood and Anxiety*, a book that aims to educate people on how exercise can help their mental health and then give them the tools to do so. It did really well at the former: the logic was compelling and the foundational evidence was well cited and mostly true (although exercise science always has wide error bars). But out of 14 people who agreed to read the book and attempt to exercise more, only three reported back to me and none of them reported an increase in exercise. So EfMaA is true and epistemically legible, but nonetheless not very useful.

True but Epistemically Illegible

[You Have About Five Words](#) is a poetic essay from Ray Arnold. The final ~paragraph is as follows:

If you want to coordinate thousands of people...

You have about five words.

This has ramifications on how complicated a coordinated effort you can attempt.

What if you need all that nuance and to coordinate thousands of people? What would it look like if the world was filled with complicated problems that required lots of people to solve?

I guess it'd look like this one.

I think the steelman of its core claim, that humans are bad at remembering long nuanced writing and the more people you are communicating with, the more you need to simplify your writing, is *obviously* true. This is good, because Ray isn't doing crap to convince me of it. He cites no evidence and gives no explanation of his logic. If I thought nuance increased with the number of readers I would have nothing to say other than "no you're wrong" or write my own post from scratch, because he gives no hooks to refute. If someone tried to argue that you get ten words rather than five, I would think they were missing the point. If I thought he had the direction right but got the magnitude of the effect wrong enough that it mattered (and he was a stranger rather than a friend), I would not know where to start the discussion.

[Ray gets a few cooperation points back by explicitly labeling this as poetry, which normally I would be extremely happy about, but it weakened its usefulness as an example for this post so right this second I'm annoyed about it.]

False but Epistemically Legible

Mindset

I think [Carol Dweck's Mindset](#) and associated work is very wrong, and I can produce [large volumes on specific points of disagreement](#). This is a sign of a work that is very epistemically legible: I know what her cruxes are, so I can say where I disagree. For all the shit I've talked about Carol Dweck over the years, I appreciate that she made it so extraordinarily easy to do so, because she was so clear on where her beliefs came from.

For example, here's a quote from *Mindset*

All children were told that they had performed well on this problem set: "Wow, you did very well on these problems. You got [number of problems] right. That's a really high score!" No matter what their actual score, all children were told that they had solved at least 80% of the problems that they answered.

Some children were praised for their ability after the initial positive feedback: "You must be smart at these problems." Some children were praised for their effort after the initial positive feedback: "You must have worked hard at these problems." The remaining children were in the control condition and received no additional feedback.

And here's [Scott Alexander's criticism](#)

This is a nothing intervention, the tiniest ghost of an intervention. The experiment had previously involved all sorts of complicated directions and tasks, I get the impression they were in the lab for at least a half hour, and the experimental intervention is changing three short words in the middle of a sentence.

And what happened? The children in the intelligence praise condition were much more likely to say at the end of the experiment that they thought intelligence was more important than effort ($p < 0.001$) than the children in the effort condition. When given the choice, 67% of the effort-condition children chose to set challenging learning-oriented goals, compared to only 8% (!) of the intelligence-condition. After a further trial in which the children were rigged to fail, children in the effort condition were much more likely to attribute their failure to not trying

hard enough, and those in the intelligence condition to not being smart enough ($p < 0.001$). Children in the intelligence condition were much less likely to persevere on a difficult task than children in the effort condition (3.2 vs. 4.5 minutes, $p < 0.001$), enjoyed the activity less ($p < 0.001$) and did worse on future non-impossible problem sets (p...you get the picture). This was repeated in a bunch of subsequent studies by the same team among white students, black students, Hispanic students...you probably still get the picture.

Scott could make those criticisms because Dweck described her experiment in detail. If she'd said "we encouraged some kids and discouraged others", there would be a lot more ambiguity.

Meanwhile, I want to criticize her for lying to children. Messing up children's feedback system creates the dependencies on adult authorities that lead to problems later in life. This is extremely bad even if it produces short-term improvements (which it doesn't). But I can only do this with confidence because she specified the intervention.

The Fate of Rome

This one is more overconfident than false. [*The Fate of Rome*](#) laid out very clearly how they were using new tools for recovering meteorological data to determine the weather 2000 years ago, and using that to analyze the Roman empire. Using this new data, it concludes that the peak of Rome was at least partially caused by a prolonged period of unusually good farming weather in the Mediterranean, and that the collapse started or was worsened when the weather began to regress to the mean.

I looked into the archeometeorology techniques and determined that they, in my judgement, had wider confidence intervals than the book indicated, which undercut the causality claims. I wish the book had been more cautious with its evidence, but I really appreciate that they laid out their reasoning so clearly, which made it really easy to look up points I might disagree with them on.

False and Epistemically Illegible

Public Health and Airborne Pathogen Transmission

I don't know exactly what the CDC's or WHO's current stance is on breathing-based transmission of covid, and I don't care, because they were so wrong for so long in such illegible ways.

When covid started, the CDC and WHO's story was that it couldn't be "airborne", because the viral particle was > 5 microns. That phrasing was already anti-legible for material aimed at the general public, because airborne has a noticeably different definition in virology ("can persist in the air indefinitely") than it does for popular use ("I can catch this through breathing"). But worse than that, they never provided any justification for the claim. This was reasonable for posters, but not everything was so space constrained, and when I looked in February 2021 I could not figure out where the belief that airborne transmission was rare was coming from. [Some researcher](#) eventually spent dozens to hundreds of hours on this and determined the 5 micron number probably came from studies of tuberculosis, which for various reasons needs to get deeper in the lungs than most pathogens and thus has stronger size

constraints. If the CDC had pointed to their sources from the start we could have determined the 5 micron limit was bullshit much more easily (the fact that many relevant people accepted it without that proof is a separate issue).

When I wrote up the Carol Dweck example, it was easy. I'm really confident in what Carol Dweck believed at the time of writing *Mindset*, so it's really easy to describe why I disagree. Writing this section on the CDC was harder, because I cannot remember exactly what the CDC said and when they said it; a lot of the message lived in implications; their statements from early 2020 are now memory holed and while I'm sure I could find them on archive.org, it's not really going to quiet the nagging fear that someone in the comments is going to pull up a different thing they said somewhere else that doesn't say exactly what I claimed they said, or that I view as of a piece with what I cited but both statements are fuzzy enough that it would be a lot of work to explain why I think the differences are immaterial....

That fear and difficulty in describing someone's beliefs is the hallmark of epistemic illegibility. The wider the confidence interval on what someone is claiming, the more work I have to do to question it.

And More...

The above was an unusually legible case of illegibility. Mostly illegible and false arguments don't feel like that. They just feel frustrating and bad and like the other person is wrong but it's too much work to demonstrate how. This is inconveniently similar to the feeling when the other person is right but you don't want to admit it. I'm going to gesture some more at illegibility here, but it's inherently an illegible concept so there will be genuinely legible (to someone) works that resemble these points, and illegible works that don't.

Marks of probable illegibility:

- The person counters every objection raised, but the counters aren't logically consistent with each other.
- You can't nail down exactly what the person actually believes. This doesn't mean they're uncertain – saying "I think this effect is somewhere between 0.1x and 10000x" is very legible, and sometimes the best you can do given the data. It's more that they imply a narrow confidence band, but the value that band surrounds moves depending on the subargument. Or they agree they're being vague but they move forward in the argument as if they were specific.
 - [Motte and Bailey](#) is a subtype of this.
- You feel like you understand the argument and excitedly tell your friends. When they ask obvious questions you have no answer or explanation.

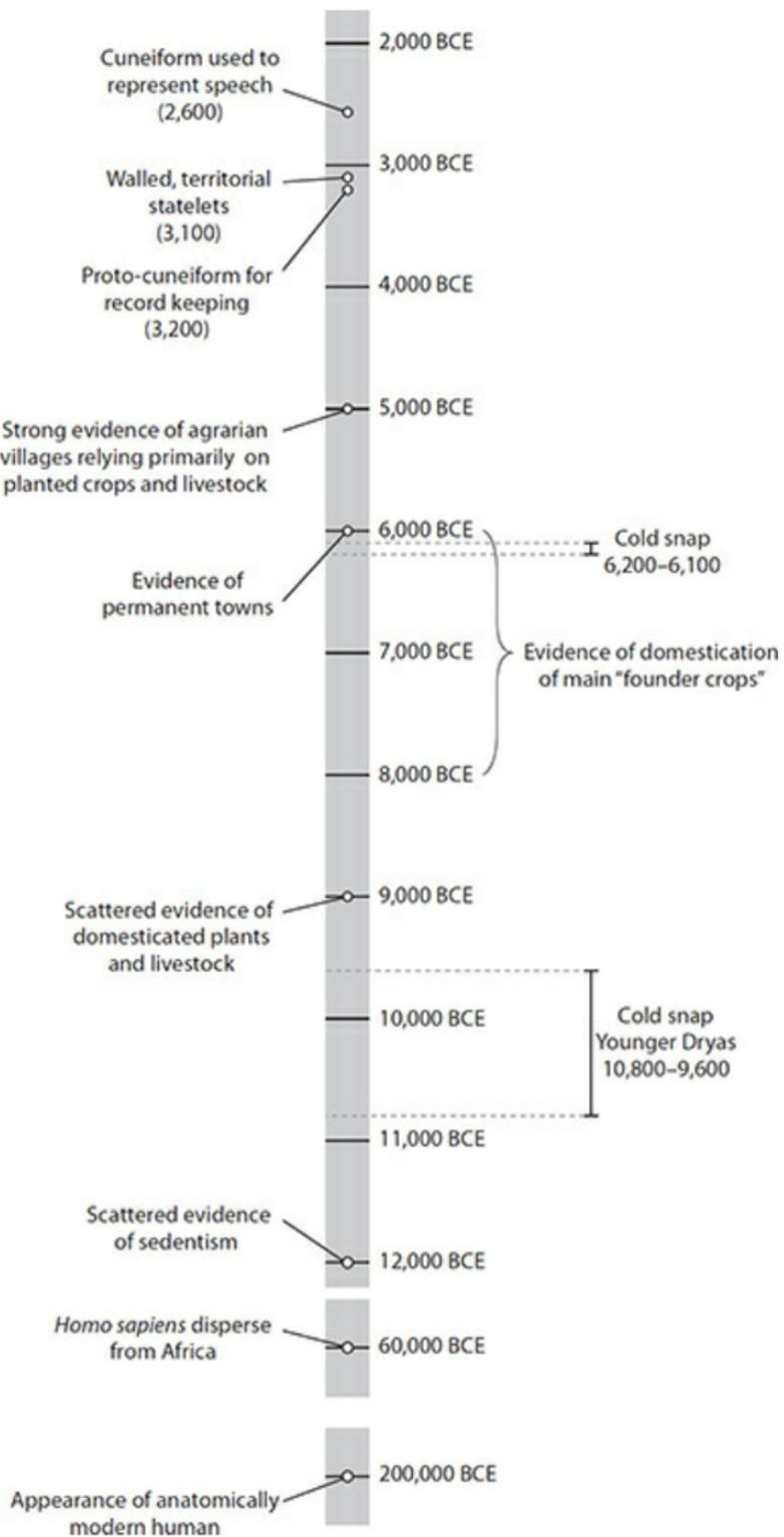
A good example of illegibly bad arguments that are specifically trying to ape legibility are a certain subset of alt-medicine advertisements. They start out very specific, with things like "there are 9804538905 neurons in your brain carrying 38923098 neurotransmitters", with rigorous citations demonstrating those numbers. Then they introduce their treatment in a way that very strongly implies it works with those 38923098 transmitters but not, like, what it does to them or why we would expect that to have a particular effect. Then they wrap it up with some vague claims about wellness, so you're left with the feeling you'll definitely feel better if you take their pill, but if you complain about any particular problem it did not fix they have plausible deniability.

[Unfortunately the FDA's rules around labeling encourage this illegibility even for products that have good arguments and evidence for efficacy on specific problems, so the fact that a product does this isn't conclusive evidence it's useless.]

Bonus Example: Against The Grain

The concept of epistemic legibility was in large part inspired by my first attempt at James C. Scott's *Against the Grain* (if that name seems familiar: Scott also coined "legibility" in the sense in which I am using it), whose thesis is that key properties of grains (as opposed to other domesticates) enabled early states. For complicated reasons I read more of AtG without epistemic checking than I usually would, and then checks were delayed indefinitely, and then covid hit, and then my freelancing business really took off... the point is, when I read *Against the Grain* in late 2019, it felt like it was going to be the easiest epistemic spot check I'd ever done. Scott was so cooperative in labeling his sources, claims, and logical conclusions. But when I finally sat down to check his work, I found serious illegibilities.

I did the spot check over Christmas this year (which required restarting the book). It was maybe 95% as good as I remembered, which is extremely high. At chapter 4 (which is halfway through the book, due to the preface and introduction), I felt kinda overloaded and started to spot check some claims (mostly factual – the logical ones all seemed to check out as I read them). A little resentfully, I checked this graph.



This should have been completely unnecessary, Scott is a decent writer and scientist who was not going to screw up basic dates. I even split the claims section of the draft into two sections, “Boring” and “Interesting”, because I obviously wasn’t going to come up with anything checking names and dates and I wanted that part to be easy to skip.

I worked from the bottom. At first, it was a little more useful than I expected – a major new interpretation of the data came out the same year the book was published, so Scott’s timing on anatomically modern humans was out of date, but not in a way that reflected poorly on him.

Finally I worked my way up to “first walled, territorial state”. Not thinking super hard, I googled “first walled city”, and got a date 3000 years before the one Scott cites. Not a big deal, he specified state, not walls. What can I google to find that out? “Earliest state”, obviously, and the [first google hit](#) does match Scott’s timing, but... what made something a state, and how can we assess those traits from archeological records? I checked, and nowhere in the preface, introduction, or first three chapters was “state” defined. No work can define every term it uses, but this is a pretty important one for a book whose full title is *Against the Grain: A Deep History of the Earliest States*.

You might wonder if “state” had a widespread definition such that it didn’t need to be defined. I think this is not the case for a few reasons. First, *Against The Grain* is aimed at a mainstream audience, and that requires defining terms even if they’re commonly known by experts. Second, even if a reader knew the common definition of what made a state, how you determine whether something was a state or merely a city from archeology records is crucial for understanding the inner gears of the book’s thesis. Third, when Scott finally gives a definition, it’s not the same as the one on wikipedia.

[longer explanation] Among these characteristics, I propose to privilege those that point to territoriality and a specialized state apparatus: walls, tax collection, and officials.

Against the Grain

States are minimally defined by anthropologist David S. Sandeford as socially stratified and bureaucratically governed societies with at least four levels of settlement hierarchy (e.g., a large capital, cities, villages, and hamlets)

Wikipedia (as of 2021-12-26)

These aren’t incompatible, but they’re very far from isomorphic. I expect that even though there’s a fairly well accepted definition of state in the relevant field(s), there are disputed edges that matter very much for this exact discussion, in which Scott views himself as pushing back against the commonly accepted narrative.

To be fair, the definition of state was not that relevant to chapters 1-3, which focus on pre-state farming. Unless, you know, your definition of “state” differs sufficiently from his.

Against The Grain was indeed very legible in other ways, but loses basically all of its accrued legibility points and more for not making even a cursory definition of a crucial term in the introduction, and for doing an insufficient job halfway through the book.

This doesn’t mean the book is useless, but it does mean it was going to be more work to extract value from than I felt like putting in on this particular topic.

Why is this Important?

First of all, it's costing me time.

I work really hard to believe true things and disbelieve false things, and people who argue illegibly make that harder, especially when people I respect treat arguments as more proven than their level of legibility allows them to be. I expect having a handle with which to say "no I don't have a concise argument about why this work is wrong, and that's a fact about the work" to be very useful.

More generally, I think there's a range of acceptable legibility levels for a given goal, but we should react differently based on which legibility level the author chose, and that arguments will be more productive if everyone involved agrees on both the legibility level and on the proper response to a given legibility level. One rule I have is that it's fine to declare something a [butterfly idea](#) and thus off limits to sharp criticism, but that inherently limits the calls to action you can make based on that idea.

Eventually I hope people will develop some general consensus around the rights and responsibilities of a given level of legibility, and that this will make arguments easier and more productive. Establishing those rules is well beyond the scope of this post.

Legibility vs Inferential Distance

You can't explain everything to everyone all of the time. Some people are not going to have the background knowledge to understand a particular essay of yours. In cases like this, legibility is defined as "the reader walks away with the understanding that they didn't understand your argument". Illegibility in this case is when they erroneously think they understand your argument. In programming terms, it's the difference between a failed function call returning a useful error message (legible), versus failing silently (illegible).

A particularly dangerous way this can occur is when you're using terms of art (meaning: words or phrases that have very specific meanings within a field) that are also common English words. You don't want someone thinking you're dismissing a medical miracle because you called it [statistically insignificant](#), or invalidating the concept of thought work because it doesn't [apply force to move an object](#).

Cruelly, misunderstanding becomes more likely the more similar the technical definition is to the English definition. I watched a friend use the term "common knowledge" to mean "everyone knows that everyone knows, and everyone knows that everyone knows..." and that metaknowledge enables actions that wouldn't be possible if it was merely true that everyone knew and thought they were the only one, and those additional possible actions are extremely relevant to our current conversation" to another friend who thought "common knowledge" meant "knowledge that is common", and had I not intervened the ensuing conversation would have been useless at best.

Costs of Legibility

The obvious ones are time and mental effort, and those should not be discounted. Given a finite amount of time, legibility on one work trades off against another piece being produced at all, and that may be the wrong call.

A second is that legibility can make things really dry. Legibility often means precision, and precision is boring, especially relative to work optimized to be emotionally activating.

Beyond that, legibility is not always desirable. For example, unilateral legibility in an adversarial environment makes you vulnerable, as you're giving people the keys to the kingdom of "effective lies to tell you".

Lastly, premature epistemic legibility kills [butterfly ideas](#), which are beautiful and precious and need to be defended until they can evolve combat skills.

How to be Legible

This could easily be multiple posts, I'm including a how-to section here more to help convey the concept of epistemic legibility than write a comprehensive guide to how to do it. The list is not a complete list, and items on it can be faked. I think a lot of legibility is downstream of something harder to describe. Nonetheless, here are a few ways to make yourself more legible, when that is your goal.

- Make it clear what you actually believe.
 - Watch out for implicit quantitative estimates ("probably", "a lot", "not very much") and make them explicit, even if you have a very wide confidence interval. The goals here are twofold: the first is to make your thought process explicit *to you*. The second is to avoid confusion - people can mean different things by "many", and I've seen some very long arguments suddenly resolve when both sides gave actual numbers.
- Make clear the evidence you are basing your beliefs on.
 - This need not mean "scientific fact" or "RCT". It could be "I experienced this a bunch in my life" or "gut feeling" or "someone I really trust told me so". Those are all valid reasons to believe things. You just need to label them.
- Make that evidence easy to verify.
 - More accessible sources are better.
 - Try to avoid paywalls and \$900 books with no digital versions.
 - If it's a large work, use page numbers or timestamps to the specific claim, removing the burden to read an entire book to check your work (but if your claim rests on a large part of the work, better to say that than artificially constrict your evidence)
 - One difficulty is when the evidence is in a pattern, and no one has rigorously collated the data that would let you demonstrate it. You can gather the data yourself, but if it takes a lot of time it may not be worth it.
 - In times past, when I wanted to refer to a belief I had in a blog post but didn't have a citation for it, I would google the belief and link to the first article that came up. I regret this. Just because an article agrees with me doesn't mean it's good, or that its reasoning is my reasoning. So one, I might be passing on a bad argument. Two, I know that, so if someone discredits the linked article it doesn't necessarily change my mind, or even create in me a feeling of obligation to investigate. I now view it as more

honest to say “I believe this but only vaguely remember the reasons why”, and if it ends up being a point of contention I can hash it out later.

- Make clear the logical steps between the evidence and your final conclusion.
- Use examples. Like, so many more examples than you think. Almost everything could benefit from more examples, especially if you make it clear when they’re skippable so people who have grokked the concept can move on.
 - It’s helpful to make clear when an example is evidence vs when it’s a clarification of your beliefs. The difference is if you’d change your mind if the point was proven false: if yes, it’s evidence. If you’d say “okay fine, but there are a million other cases where the principle holds”, it’s an example. One of the mistakes I made with early epistemic spot checks was putting too much emphasis on disproving examples that weren’t actually evidence.
- Decide on an audience and tailor your vocabulary to them.
 - All fields have words that mean something different in the field than in general conversation, like “[work](#)”, “[airborne](#)”, and “[significant](#)”. If you’re writing within the field, using those terms helps with legibility by conveying a specific idea very quickly. If you’re communicating outside the field, using such terms without definition hinders legibility, as laypeople misapply their general knowledge of the English language to your term of art and predictably get it wrong. You can help on the margins by defining the term in your text, but I consider some uses of this iffy.
 - The closer the technical definition of a term is to its common usage, the more likely this is to be a problem because it makes it much easier for the reader to think they understand your meaning when they don’t.
 - At first I wanted to yell at people who use terms of art in work aimed at the general population, but sometimes it’s unintentional, and sometimes it’s a domain expert who’s bad at public speaking and has been unexpectedly thrust onto a larger stage, and we could use more of the latter, so I don’t want to punish people too much here. But if you’re, say, a journalist who writes a general populace book but uses an academic term of art in a way that will predictably be misinterpreted, you have no such excuse and will go to legibility jail.
 - A skill really good interviewers bring to the table is recognizing terms of art that are liable to confuse people and prompting domain experts to explain them.
- Write things down, or at least write down your sources. I realize this is partially generational and Gen Z is more likely to find audio/video more accessible than written work, and accessibility is part of legibility. But if you’re relying on a large evidence base it’s very disruptive to include it in audio and very illegible to leave it out entirely, so write it down.
- Follow all the rules of normal readability – grammar, paragraph breaks, no run-on sentences, etc.

A related but distinct skill is making your own thought process legible. John Wentworth describes that [here](#).

Synthesis

“This isn’t very epistemically legible to me” is a valid description (when true), and a valid reason not to engage. It is not automatically a criticism.

“This idea is in its butterfly stage”, “I’m prioritizing other virtues” or “this wasn’t aimed at you” are all valid defenses against accusations of illegibility as a criticism (when true), but do not render the idea more legible.

“This call to action isn’t sufficiently epistemically legible to the people it’s aimed at” is an *extremely* valid criticism (when true), and we should be making it more often.

I apologize to Carol Dweck for 70% of the vigor of my criticism of her work; she deserves more credit than I gave her for making it so easy to do that. I still think she’s wrong, though.

Epilogue: Developing a Standard for Legibility

As mentioned above, I think the major value add from the concept of legibility is that it lets us talk about whether a given work is sufficiently legible for its goal. To do this, we need to have some common standards for how much legibility a given goal demands. My thoughts on this are much less developed and by definition common standards need to be developed by the community that holds them, not imposed by a random blogger, so I’ll save my ideas for a different post.

Epilogue 2: Epistemic Cooperation

Epistemic legibility is part of a broader set of skills/traits I want to call epistemic cooperation. Unfortunately, legibility is the only one I have a really firm handle on right now (to the point I originally conflated the concepts, until a few conversations highlighted the distinction- thanks friends!). I think epistemic cooperation, in the sense of “makes it easy for us to work together to figure out the truth” is a useful concept in its own right, and hope to write more about it as I get additional handles. In the meantime, there are a few things I want to highlight as increasing or signalling cooperation in general but not legibility in particular:

- Highlight ways your evidence is weak, related things you don’t believe, etc.
- Volunteer biases you might have.
- Provide reasons people might disagree with you.
- Don’t emotionally charge an argument beyond what’s inherent in the topic, but don’t suppress emotion below what’s inherent in the topic either.
- Don’t tie up brain space with data that doesn’t matter.

Thanks to Ray Arnold, John Salvatier, John Wentworth, and Matthew Graves for discussion on this post.

12 interesting things I learned studying the discovery of nature's laws

I've been thinking about whether I can discover laws of agency and wield them to prevent AI ruin (perhaps by building an AGI myself in a different paradigm than machine learning).

So far I've looked into the history of the discovery of physical laws (gravity in particular) and mathematical laws (probability theory in particular). Here are 12 things I've learned or been surprised by.

1.

Data-gathering was a crucial step in discovering both gravity and probability theory. One rich dude had a whole island and set it up to have lenses on lots of parts of it, and for like a year he'd go around each day and note down the positions of the stars. Then this data was worked on by others who turned it into equations of motion.

2.

Relatedly, looking at the celestial bodies was a big deal. It was almost the whole game in gravity, but also a little helpful for probability theory (specifically the normal distribution was developed in part by noting that systematic errors in celestial measuring equipment followed a simple distribution).

It hadn't struck me before, but putting a ton of geometry problems on the ceiling for the entire civilization led a lot of people to try to answer questions about it. (It makes Eliezer's choice in [That Alien Message](#) apt.) I'm tempted in a munchkin way to find other ways to do this, like to write a math problem on the surface of the moon, or petition Google to put a prediction market on its home page, or something more elegant than those two.

3.

Probability theory was substantially developed around real-world problems! I thought math was all magical and ivory tower, but it was much more grounded than I expected.

After a few small things like accounting and insurance and doing permutations of the alphabet, games of chance (gambling) was what really kicked it off, with Fermat and Pascal trying to figure out the expected value of games (they didn't phrase it like that, they put it more like "if the game has to stop before it's concluded, how should the winnings be split between the players?").

Other people who consulted with gamblers also would write down data about things like how often different winning hands would come up in different games, and discovered simple distributions, then tried to put equations to them. Later it was developed further by people trying to reason about gases and temperatures, and then again in understanding clinical trials or large repeated biological experiments.

Often people discovered more in this combination of “looking directly at nature” and “being the sort of person who was interested in developing a formal calculus to model what was going on”.

4.

Thought experiments about the world were a big deal too! Thomas Bayes did most of his math this way. He had a thought experiment that went something like this: his assistant would throw a ball on a table that Thomas wasn’t looking at. Then his assistant would throw more balls on the table, each time saying whether it ended up to the right or the left of the original ball. He had this sense that each time he was told the next left-or-right, he should be able to give a new probability that the ball was in any particular given region. He used this thought experiment a lot when coming up with Bayes’ theorem.

5.

Lots of people involved were full-time inventors, rich people who did serious study into a lot of different areas, including mathematics. This is a weird class to me. (I don’t know people like this today. And most scientific things are very institutionalized, or failing that, embedded within business.)

Here’s a quote I enjoyed from one of Pascal’s letters to Fermat when they founded the theory of probability. (For context: de Mere was the gambler who asked Pascal for help with a confusion he had.)

“I have not time to send you the demonstration of a difficulty which greatly astonished M. de Mere, for he has a very good mind, but he is not a geometer (this is, as you know, a great defect)...” - Blaise Pascal

6.

In Laplace’s seminal work putting probability theory on a formal footing, he has a historical section at the end praising all the people who did work, how great they were and how beautiful their work was. Then he has one line on Bayes where he calls his work “a little perplexing”.

*“Bayes, in the *Transactions Philosophiques* of the year 1763, sought directly the probability that the possibilities indicated by past experiences are comprised within given limits; and he has arrived at this in a refined and very ingenious manner, although a little perplexing.”*

Also, whenever you feel like you’ve missed out on your glorious youth, note that Thomas Bayes got interested in probability theory in his 50s, and died aged 59. He was not formally trained in math in his youth.

7.

I watched a talk by Pearl about his causal models, and I was struck by the extent to which he had a “philosophy” of counterfactual inference. It had seemed pretty possible to me he would have said “here was a problem, and here is my solution”, but instead he had a lot to say about counterfactuals and how he thought about them conceptually that wasn’t in the math.

It reminds me of my impression that Daniel Kahneman (and Amos Tversky) have strong models of how their minds work, of which the heuristics & biases literature is a legitimized component of, but certainly does not capture the whole thing.

Relatedly, in a lecture by Feynman on seeking new laws, he says that some people say “don’t talk about what you cannot measure”. He says he agrees insofar as your theories need measurable predictions, but he doesn’t agree that people should stop discussing their whole philosophies, as the philosophies seem to help some people come up with good guesses about laws.

I think in the past I could have found myself unable to justify my interest in the philosophy of something as more than a personal interest. Now I have a practical justification, which is that it helps me come up with guesses about how nature works! And my current guess is that many people who were successful at that had unique and well-developed philosophies.

8.

Pearl himself says that he has discovered two laws, and once you have them, you can fire him, because the rest is just algebra! And he calls it a calculus of counterfactuals, just like Newton and Bayes and everyone did. Fascinating.

I couldn’t find anything on what problems Pearl was thinking about when he came up with his calculus of counterfactuals. Like, was he personally trying to analyze clinical trials? Was he a mathematician who was friends with people doing large experiments and thought the math was interesting? I want to know what part of the world he was in contact with when developing it.

9.

I updated against expecting to resolve scientific disagreements at the time when the correct theory is known. Let me explain.

In the discovery of gravity, there were a lot of anomalies that didn’t fit the data. For instance, Jupiter didn’t follow the law: its orbit was a more elongated ellipse when it was further away. Uranus’s orbit would jiggle a bit sometimes. Also there were two stars who didn’t orbit their collective center of gravity, but instead some other point within the ellipse. At this point I would have been like “yeah, nice try, but your theory isn’t fitting the details”.

Want to know what they said at the time? (Spoilers ahead.) For the stars, they said that we were probably just looking at them *at a funny angle* and that’s why it didn’t work. For Uranus, they said there was *an invisible planet* that was knocking it off-course. And for Jupiter, they said the *light was moving too slowly* for the measurements to work out.

To me this seems like an awful lot of complexity cost weighing on a theory. Now it’s no longer just a theory, it’s also a lot of explaining exceptions with unlikely stories. The star-angle one doesn’t even seem testable, it gives me a Scott-Alexander-like sense of *this explanation gives me so many degrees of freedom that I can probably explain away loads more anomalies with it*.

Anyway... they were all right.

From Uranus's wobbles, they found Neptune. The stars were indeed rotated at an angle. And they did some experiments and found out that light did have a speed and this explained the Jupiter issue, and opened up a whole new area of inquiry about light.

Very impressive in retrospect, but I feel like I couldn't have gotten this right at the time.

My update is further in the direction that Jacob's post [The Copernican Revolution from the Inside](#) argues for, which is that if two different people had different theories at the time, I do not anticipate the disagreement being able to be "clearly resolvable" at all, and do expect for it to involve a great number of judgment calls, in large part dependent on one's "philosophy" of how to make those calls in this domain.

10.

Feynman has a wonderful quote on the art of guessing nature's laws that includes at least two paths not discussed above. That said I don't understand them, in particular the ways that quantum mechanics was discovered. (I'm tempted to dig into that some.)

I've put the full quote in this footnote^[1], recommended.

11.

One confusion I wrote down in advance was "I still don't quite know how to predict that there will not be a simple mathematical apparatus that explains something. Why the motion of the planets, why the game of chance, why not the color of houses in England or the number of hairs on a man's head?"

Looking back on this, I don't know whether I got a direct answer, but I now feel that my answer is something like "look for the places where Nature will show herself directly". Obviously that's not a very well-specified answer, but I feel like it points to a real distinction.

12.

I also made an advance prediction: "I guess I also make the advance prediction that most of the rest of the [probability] math was developed by people who liked symbol manipulation more than people doing real-world problem solving. But I would be interested to be surprised here."

This prediction was false! It took both! All the probability math was developed by people who liked using math to reason rigorously about the world, and who were interested in understanding the real world! There were exceptions like Bayes who relied a great deal on thought-experiment, though sort of still "about" the world, not just about symbols.

When I thought of math previously I thought about my math friends in academia, who just sort of entered the abstract world as a starting point and lived in there. ("My professor does work in flat-spherical-manifold-density-vector-spaces, so I'm trying to prove something there too!") Now I think of people trying to reason about particular parts of the world I live in, and who are trying to make an externalized symbolic calculus that can do that reasoning for them.

Next Step

The natural next step of my investigation is to learn more about how key discoveries in areas like optimization and information theory and game theory were made. How did nature show herself to these discoverers? I have written down a few advance predictions for if I continue seeking this information...

1. [^](#)

Feynman, on the art of guessing nature's laws, in his final lecture for BBC's Messenger Lectures:

"Or look at history, you first start out with Newton: he [was] in a situation where he had incomplete knowledge, and he was able to get the laws by putting together ideas which all were relatively close to experiment—there wasn't a great distance between the observations and the test."

"Now, the next guy who did something—another man who did something great—was Maxwell, who obtained the laws of electricity and magnetism. But what he did was this, he put together all the laws of electricity due to Faraday and other people that came before him, and he looked at them and he realized that they were mutually inconsistent; they were mathematically inconsistent. In order to straighten it out he had to add one term to an equation."

"By the way, he did this by inventing a model for himself of idler wheels, and gears, and so on, in space. Then he found what the new law was, and nobody paid much attention, because they didn't believe in the idler wheels. We don't believe in the idler wheels today, but the equations that he obtained were correct. So the logic may be wrong, but the answer is all right."

"In the case of relativity, the discovery of relativity was completely different: there was an accumulation of paradoxes; the known laws gave inconsistent results, and it was a new kind of thinking, a thinking in terms of discussing the possible symmetries of laws. It was especially difficult because it was for the first time realized how long something like Newton's laws could be right—and still ultimately be wrong—and, second, that ordinary ideas of time and space that seem so instinctive could be wrong."

"Quantum mechanics was discovered in two independent ways, which is a lesson. There, again, and even more so, an enormous number of paradoxes were discovered experimentally, things that absolutely couldn't be explained in any way by what was known—not that the knowledge was incomplete, but the knowledge was too complete!: your prediction was, this should happen; it didn't."

The two different routes were: one, by Schrodinger, who guessed the equations; another, by Heisenberg, who argued that you must analyze what's measurable. So two different philosophical methods reduced to the same discovery in the end."

"More recently, the discovery of the laws of this [weak decay] interaction, which are still only partly known, add quite a somewhat different situation: this time it was a case of incomplete knowledge, and only the equation was guessed. The special difficulty this time was that the experiments were all wrong—all the experiments were wrong."

"Now, how can you guess the right answer when, when you calculate the results it disagrees with the experiment, and you have the courage to say the experiments must be wrong. I'll explain where the courage comes from in a minute."

"Now, I'm sure that history does not repeat itself in physics, as you see from this list, and the reason is this: any scheme—like, "Think of symmetry laws," or "Put the equations in mathematical form," or any of these schemes "Guess equations," and so on—are known to everybody now, and they're tried all the time. So if the place where you get stuck is not that—and you try that right away: we try looking for symmetries; we try all the things that have been tried before, but we're stuck—so it must be another way next time.

Each time that we get in this log jam of too many problems, it's because the methods that we're using are just like the ones we used before. We try all that right away, but the new discovery is going to be made in a completely different way—so history doesn't help us very much."

[Beta Feature] Google-Docs-like editing for LessWrong posts

TL;DR: LessWrong now has similar features to Google Docs. Warning! Still rough around the edges. To enable the collaborative editor, you must check "Opt into experimental features" in your [account settings](#) and then press the green "Share" button that appears when editing your post.

You can experiment with commenting and suggesting on [this post with this link](#).

It's been a loooong time coming^[1] but at last, we are ready to unveil collaborative editing features for the LessWrong text editor. These features will be familiar to those used to working in Google Docs:

- Multiple users can edit a document at once
- Fine-grained permissions for viewing/commenting/editing by link or username
- Inline comments (**only viewable while in edit mode**)
- Making and accepting suggested edits
- Automatic saving
- Version history viewer

Some advantages of using LW-Docs with collaborative editing:

- **LW-Docs supports LaTeX, unlike Google Docs.**
- If you use entirely LW-Docs, you won't have broken footnotes, unlike with copying from G-Docs.
- While writing your post, you'll know what the end result will look like (same font and line width) which helps you optimize paragraphs and layout for looking good when published.
- You can continue to get inline feedback and suggestions on your post even after you've copied it over to LessWrong.
 - These can then be seamlessly integrated into your live post.

Why LessWrong

SHARE
(BETA)

□ Alignment Forum 

I connected user (me) Ruby

Commenting ▾ SAVED

Collective Thought

¶ Humans are collective thinkers. Our progress mapping out the world and all its mysteries has only ever been made because one human got to add their thoughts to another's. That we think independently is but an illusion, derived from the fact that we have discrete brains separated from one another by skin and bone.

Francis Bacon was not the first to opine on scientific methodology, Maxwell was not the first to play with electricity, Euler was not the first geometer, nor Newton the first physicist. Each of those got to build upon the ideas of others—and thanks to technology—the thoughts of those they never even met.

The technology is crucial. Humanity did not always comprise a machine that could figure out things that no single human ever could. There's been a steady stream of upgrades for millennia: our vocal chords changed to increase the range of speech we could make^[1]; our brains developed specialized regions for speech comprehension and production^{[2][3]}; we invented clay tablets, vellum, ink, the printing press, schools, lectures, apprenticeship, textbooks, journals, conferences, newspapers, typewriters, computers, search engines, and a globe-spanning network that beams information at the speed of light. Needless to say, this list woefully incomplete.

Each of these innovations made us more powerful. More than ever before, humans can pool their thoughts to hack away at hard, important, or interesting problems.

We're not finished

Yet why should we imagine that we are done? If we humans have improved our capacity

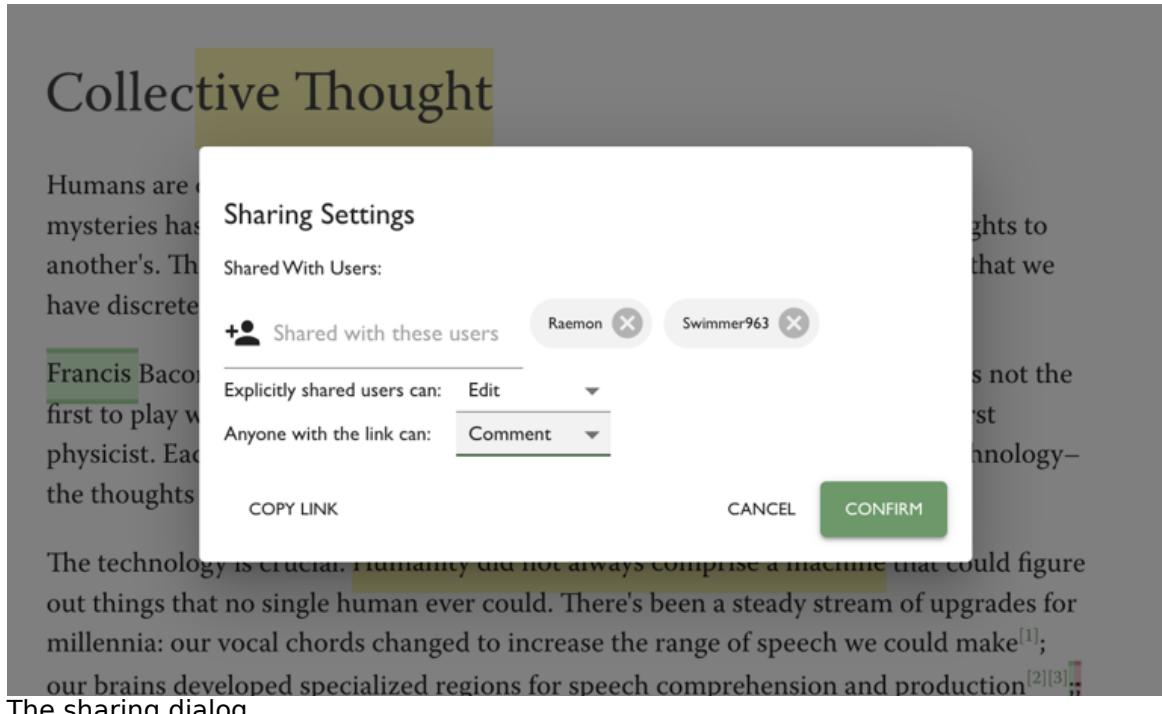
INLINE COMMENTS AND SUGGESTIONS ARE ONLY VIEWABLE IN EDIT MODE TO THOSE WITH PERMISSIONS

Ruby Today 04:24PM ✓ ✗
Replace: "an" with "an"

Ruby Today 04:24PM ✓ ✗
Insert: "only "
Reply...

Ruby Today 04:26PM ✓ ✗
Insert: "Francis "

Swimmer963 02-03-2022 02:05PM 📁
"humanity" feels slightly weird here, I would maybe say "civilization"
Ruby 02-03-2022 02:15PM 📁
Why does humanity feel weird? I think I more strongly mean "all the humans" than "civilization" which somehow connotes more building and stuff.



The sharing dialog

How to enable collaborative editing on your post

Step 1: Opt in to experimental features in your [account settings](#).

- Hide the tagging progress bar
- Hide the frontpage book ad
- Hide Intercom
- Opt into experimental features
- Activate Markdown Editor
- Restore the previous WYSIWYG editor
- Hide other users' Elicit predictions until I have predicted myself

Step 2: While editing your post, click the green "Share" button in the top right. Enable permissions for some other users. Boom! Your post is now in **collaborative mode**.

[Beta Feature] Google-Docs-like editing for LessWrong posts

SHARE
(BETA)

Step 3: Users explicitly shared on your post will receive a notification. Send the url to anyone else you want to grant access.

Step 4: When in **collaborative mode**, the text of your post is automatically saved. To view how your post will look when published, press the **Preview** button. To make the current state of your document live, press the **Publish** button.

GET FEEDBACK

PREVIEW

PUBLISH

The collaborative editor allows you (and others) to continue editing and commenting on a post even once it's been published. Edits aren't automatically published to the live version. To update the published version to the current state of the document in editing, press "Publish".

Step 5: Leave comments and suggestions.

To leave a comment, highlight text and press the "Comment" button



This feature is still under development. Any feedback from early-adopters is hugely helpful. Just leave a comment on this post or message us on Intercom.

To enable track changes, use the track changes button in the popup menu (third icon from the right) or set your Mode to Commenting in the header bar.



The header bar. Being in Comment mode means you only suggest changes rather than actually make them.

Step 6: Leave feedback about the feature!

This feature is still under development. Any feedback from early-adopters is hugely helpful. Just leave a comment on this post or message us on Intercom.

Warning! Rough around the edges

We're releasing this feature in beta mode because it still requires a few more finishing touches and might be a little confusing to use in places. As above, any feedback is greatly appreciated.

Why collaborative editing?

When developing features, there's always a question of "does anyone want this?" In the case of collaborative editing features, there's good evidence of demand from Google Docs.

There's a common workflow that goes: (1) write a draft in Google Docs, (2) invite some close friends or collaborators to give feedback, (3) incorporate feedback, (4) copy to LessWrong, (5) publish.

Drawbacks of this workflow are (a) overhead of copying and reformatting the post, (b) enforcing a hard break between the feedback stage and the publication stage, (c) Google Docs does not support LaTeX, (d) valuable comments left in the feedback stage never get published to the wider world.

By introducing collaborative editing to LessWrong, we address (a), (b), and (c). We haven't yet made it so in-line comments in editing mode can be published in the final version, but we'll look into ways to allow for that. I also expect that we'll add additional features^[2] to our editor that Google Docs doesn't have, such that users will benefit from the possibility of doing all their writing on LessWrong.

Having collaborative editing features on LessWrong also lets us start to build programs that rely on easy ways to give people feedback on their drafts. For example, I'd like to run a writing and research workshop for students that involves peer and mentor feedback. With collaborative editing on LessWrong, that will now be much more convenient to do.

What's your writing workflow?

If you're an author, I'd love to hear how collaborative editing features do or don't help you with your workflow, or what you'd really like to see us build. Feel free to comment on this post or message us on Intercom.

Thanks and good luck!

1. ^

It's been ~18 months since the first steps towards this were taken.

2. ^

One feature I'm particularly excited about is "link-searching". In the same way that one can @-mention people on Facebook, I'd like to make it so you can easily link to post and wiki-tags by typing @ or # and then using a few letters to search for the resource you want to link to.

IMO challenge bet with Eliezer

Eliezer and I publicly stated some predictions about AI performance on the IMO by 2025. In honor of OpenAI's post [Solving \(Some\) Formal Math Problems](#), it seems good to publicly state and clarify our predictions, have a final chance to adjust them, and say a bit in advance about how we'd update.

The predictions

Eliezer and I had [an exchange in November 2021](#).^[1] My final prediction (after significantly revising my guesses after looking up IMO questions and medal thresholds) was:

I'd put 4% on "For the 2022, 2023, 2024, or 2025 IMO an AI built before the IMO is able to solve the single hardest problem" where "hardest problem" = "usually problem #6, but use problem #3 instead if either: (i) problem 6 is geo or (ii) problem 3 is combinatorics and problem 6 is algebra." (Would prefer just pick the hardest problem after seeing the test but seems better to commit to a procedure.)

Maybe I'll go 8% on "gets gold" instead of "solves hardest problem."

Eliezer spent less time revising his prediction, but said (earlier in the discussion):

My probability is *at least* 16% [on the IMO grand challenge falling], though I'd have to think more and Look into Things, and maybe ask for such sad little metrics as are available before I was confident saying how much more. Paul?

EDIT: I see they want to demand that the AI be open-sourced publicly before the first day of the IMO, which unfortunately sounds like the sort of foolish little real-world obstacle which can prevent a proposition like this from being judged true even where the technical capability exists. I'll stand by a >16% probability of the technical capability existing by end of 2025

So I think we have Paul at <8%, Eliezer at >16% for AI made before the IMO is able to get a gold (under time controls etc. of grand challenge) in one of 2022-2025.

Separately, we have Paul at <4% of an AI able to solve the "hardest" problem under the same conditions.

I don't plan to revise my predictions further, but I'd be happy if Eliezer wants to do so any time over the next few weeks.

Earlier in the thread I clarified that my predictions are specifically about gold medals (and become even sharper as we move to harder problems), I am not surprised by silver or bronze. My guess would be that Eliezer has a more broad distribution. The comments would be a good place for Eliezer to state other predictions, or take a final chance to revise the main prediction.

How I'd update

The informative:

- I think the IMO challenge would be significant direct evidence that powerful AI would be sooner, or at least would be technologically possible sooner. I think this would be fairly significant evidence, perhaps pushing my 2040 TAI probability up from 25% to 40% or something like that.
- I think this would be significant evidence that takeoff will be limited by sociological facts and engineering effort rather than a slow march of smooth ML scaling. Maybe I'd move from a 30% chance of hard takeoff to a 50% chance of hard takeoff.
- If Eliezer wins, he gets 1 bit of epistemic credit.^{[2][3]} These kinds of updates are slow going, and it would be better if we had a bigger portfolio of bets, but I'll take what we can get.
- This would be some update for Eliezer's view that "the future is hard to predict." I think we have clear enough pictures of the future that we have the right to be surprised by an IMO challenge win; if I'm wrong about that then it's general evidence my error bars are too narrow.

The uninformative:

- This is mostly just a brute test of a particular intuition I have about a field I haven't ever worked in. It's still interesting (see above), but it doesn't bear that much on deep facts about intelligence (my sense is that Eliezer and I are optimistic about similar methods for theorem proving), or heuristics about trend extrapolation (since we have ~no trend to extrapolate), or on progress being continuous in crowded areas (since theorem proving investment has historically been low), or on lots of pre-singularity investment in economically important areas (since theorem proving is relatively low-impact). I think there are lots of other questions that *do* bear on these things, but we weren't able to pick out a disagreement on any of them.

If an AI wins a gold on some but not all of those years, without being able to solve the hardest problems, then my update will be somewhat more limited but in the same direction. If an AI wins a bronze/silver medal, I'm not making any of these updates and don't think Eliezer gets any credit unless he wants to stake some predictions on those lower bars (I consider them much more likely, maybe 20% for "bronze or silver" vs 8% on "gold," but that's less well-considered than the bets above, but I haven't thought about that at all).

1. [▲]

We also looked for claims that Eliezer thought were very unlikely, so that he'd also have an opportunity to make some extremely surprising predictions. But we weren't able to find any clean disagreements that would resolve before the end of days.

2. [▲]

I previously added the text: "So e.g. if Eliezer and I used to get equal weight in a mixture of experts, now Eliezer should get 2x my weight. Conversely, if I win then I should get 1.1x his weight." But I think that really depends on how you want to assign weights. That's a very natural algorithm that I endorse generally, but given that neither of us really has thought carefully about this question it would be reasonable to just not update much one way or the other.

3. ^

More if he chooses to revise his prediction up from 16%, or if he wants to make a bet about the "hardest problem" claim where I'm at 4%.

Intro to Naturalism: Orientation

A note on how to approach this sequence:

If you were exactly like me, I would ask you to savor this sequence, not scarf it. I would ask you to approach each of these essays in an expansive, lingering, thoughtful sort of mood. I would ask you to read them a little bit at a time, perhaps from a comfortable chair with a warm drink beside you, and to take breaks to make dinner, sing in the car, talk to your friends, and sleep.

These essays are reflections on the central principles I have gradually excavated from my past ten years of intellectual labor. I am a very slow thinker myself; if you move too quickly, I expect we'll miss each other completely.

There's a certain kind of thing that happens when a person moves quickly, and relies a lot on their built-up structures—their familiar, tried-and-true habits of thought and perception. There is a *different* kind of thing that happens when a person can step back and bring those very structures into view, rather than standing atop them. I'm hoping for the latter.

But since you're *not* exactly like me, there might be a better way to approach this sequence, in your particular case, than the exact one I'd suggest to myself. I hope you'll take a moment to check.

What matters to me is not how fast you read, or how many sittings it takes; what matters is that you create for yourself enough space to explore, to observe the real world beyond all these words, to watch how your own thoughts and experiences unfold in dialog with mine. Any method that allows you to maintain that kind of space as you read is perfect, as far as I'm concerned.

"Naturalism" is a label for a conceptual framework, investigatory discipline, and semi-formalized way of looking at and learning about the world. I've been developing and teaching naturalism for the past couple of years, if you start counting on the day I chose the term, or since 2013, if you take a more historical perspective. I've made some [relevant content available](#), but I've had trouble writing a straightforward introductory post.

The reason for this, as far as I can tell, is that the naturalist perspective is suspicious of categories, projections, and preconceptions, and seeks to move closer toward (relatively) unfiltered, direct observations. It's specifically a frame-breaking and frame-escaping discipline, so it's hard to describe in frame-terms without being importantly misleading.

I ardently desire not to mislead anyone.

There's a saying I like a lot, which goes: "A man with one watch knows what time it is; a man with two is never sure."

(When I first heard this, I needed to pause for a moment, to let it sink in. It helped me to actually visualize wearing a watch on each wrist, then checking the time.)

The reason I like this saying is that it reminds me to be confused, in an appropriate fashion. "Confused" might even be too weak of a word—it's almost like it reminds me to be *scared*, in an appropriate fashion.

I mean, sure—for most things, I don't have to know what time it *actually* is, with sufficient precision that the off-ness of my watch makes a meaningful difference. The claim here is not that absolute clarity is required at all times.

But there is indeed an unfortunate property of having-a-watch, which is that it provides me with an *answer* to the question “what time is it?”

It provides that answer clearly, and specifically, and unambiguously. It provides that answer with *more confidence* than it ought to, like a calculation that doesn’t attend to significant digits. And if I’m not careful, then with my watch right in front of me, it’s very easy to *lose track* of the fact that I do not, in fact, know exactly what time it is. To forget that what I really know is what time it *almost* is.

This is what our concepts do for us. They are usually a strict upgrade over “entirely too much information for us to even begin to process or handle”; but if you lean on them too heavily, or too unthinkingly, they become actively misleading. Actively *harmful*, in cases where precision and accuracy genuinely matter, and being subtly wrong is disastrous.

And concepts *encourage* us to lean. They’re sturdy! Sensible! Comforting! They soothe confusion, make the world seem more predictable and comprehensible, give us the surface sensation of control (or at least understanding). It’s nice to have *answers*.

But the map is not the territory.

It’s easy to look up at the sky, and name the constellations, without losing track of your knowledge that there isn’t really a Great Bear up there. We know that the constellations aren’t “real,” that they’re just there to help us chunk and cluster and orient and discuss.



But constellations are an unusually transparent construction. In the set of fake concepts that we impose on messy reality, they're unusually candid about their fakeness. Their arbitrary nature is kind enough to be apparent and obvious.

Many concepts are much less wearing-their-fakeness-on-their-sleeve. Constellations don't bear all that much resemblance to actual stars, so it's easy to avoid getting confused. But a lot of concepts really look quite similar to the thing they're modeling, and are therefore much more seductive, mesmerizing, convincing, befuddling. Much more in-the-way, much more likely to distract, much harder to set aside and see past.

The concept Harry's mind had of the rubber eraser as a single object was *obvious nonsense*.

It was a map that didn't and *couldn't* match the territory.

Human beings modeled the world using stratified levels of organization, they had *separate thoughts* about how countries worked, how people worked, how organs worked, how cells worked, how molecules worked, how quarks worked.

When Harry's brain needed to think about the eraser, it would think about the rules that governed erasers, like "erasers can get rid of pencil-marks". Only if Harry's brain needed to predict what would happen on the lower chemical level, only then would Harry's brain start thinking - as though it were a separate fact - about rubber molecules.

But that was all in the *mind*.

Harry's mind might have separate *beliefs* about rules that governed erasers, but there was no *separate law of physics* that governed erasers.

Harry's mind modeled reality using multiple levels of organization, with different beliefs about each level. But that was all in the *map*, the true territory wasn't like that, *reality itself* had only a *single* level of organization, the quarks, it was a unified low-level process obeying mathematically simple rules.

It is *genuinely difficult* to notice that an eraser is something other than "an eraser"—to circumvent the well-intentioned shortcircuiting that our brains are so practiced at doing.

And to be clear: it's usually not necessary to notice that the mental category "eraser" is glossing over a bunch of detail. It usually does not matter; our concepts are ubiquitous in large part because they tend to be sufficient, adequate for our purposes.

But there are times when it's absolutely crucial to be un-hypnotized, when it's absolutely crucial to be aware of the difference between [what's happening] and [the layer of interpretation we've draped like a blanket over what's happening].

And there's something frightening (to me, at least) about the idea of such a crucial moment arising and people *not noticing it*, because they *aren't even aware that they're draping a blanket*. Or noticing that they need to set aside the blanket, but not knowing how to actually do so.

Which is why I've devoted so many of my resources to developing naturalism. It's an important facet of mature rationalist practice, and it's mostly missing from our collective toolkit.

Notice, though, that "naturalism" is *itself a concept*. It's a constellation painted somewhat arbitrarily over a multidimensional cluster of phenomena, pretending to be real. It's easy to say that X is a part of naturalism and Y is not, and to forget that there just *isn't any boundary* out there in the territory.



But in order to properly draw your attention to the cluster, I think I sort of have to paint those lines. Human brains (mine included) have a really hard time getting excited about vast collections of vaguely adjacent points; in order to produce something useful and comprehensible, I have to pretend that there's a Thing, there.

I think doing so is instrumentally useful, and I think that (when done honestly, as this intro sequence is attempting to do) it's not actually misleading, or self-undermining. This is a fundamental thesis of naturalism: that there are points, and there are paintings we superimpose upon them, and that *these things are different*. That the constellations are of a wholly different nature than the stars.

Doesn't mean we don't need the conceptual overlay. We just want to know, in any given moment, whether we're dealing more with paintings, or more with the things they're meant to depict.

The constellation I will paint in this sequence is a single sentence. It's a sentence I built one word at a time, sketched atop a cluster of five stars I've picked out from my view of the night sky.

The sentence is a summary of naturalism after-the-fact. It will do almost nothing to help you understand the stars themselves, the real thing that I try to do with my mind day in and day out.

But it may serve to guide your attention to those stars. It may prompt you to look more closely, for yourself, at the reality hidden behind the tidy painting.

The sentence, which I will discuss piece by piece throughout my introductory sequence, is this:

Knowing the territory takes patient and direct observation.

The sentence forms the outline of my sequence, more or less:

- Knowing
- The Territory
- Observation
- Patient Observation
- Direct Observation

My only goal in this sequence is to communicate what I mean by the sentence, "Knowing the territory takes patient and direct observation."

Here is what will happen in this sequence: I will pick out the concepts that seem central to my understanding of naturalism; I will name them with words; and I will do my best to tell you what I mean by those words.

That is all.

There are a few things you might expect from an introductory sequence that I will not even try to accomplish. I want to be clear about my intentions.

I will not try to argue for the truth of the proposition the sentence picks out. It's true, I think, that knowing the territory takes patient and direct observation. But I won't try to convince you of that here.

I won't tell you what would change my mind, or what I'd expect to see if I were wrong. I won't tell you how I think you could find out if I were correct, or if I were not. I will not present evidence. I will not engage with counterarguments.

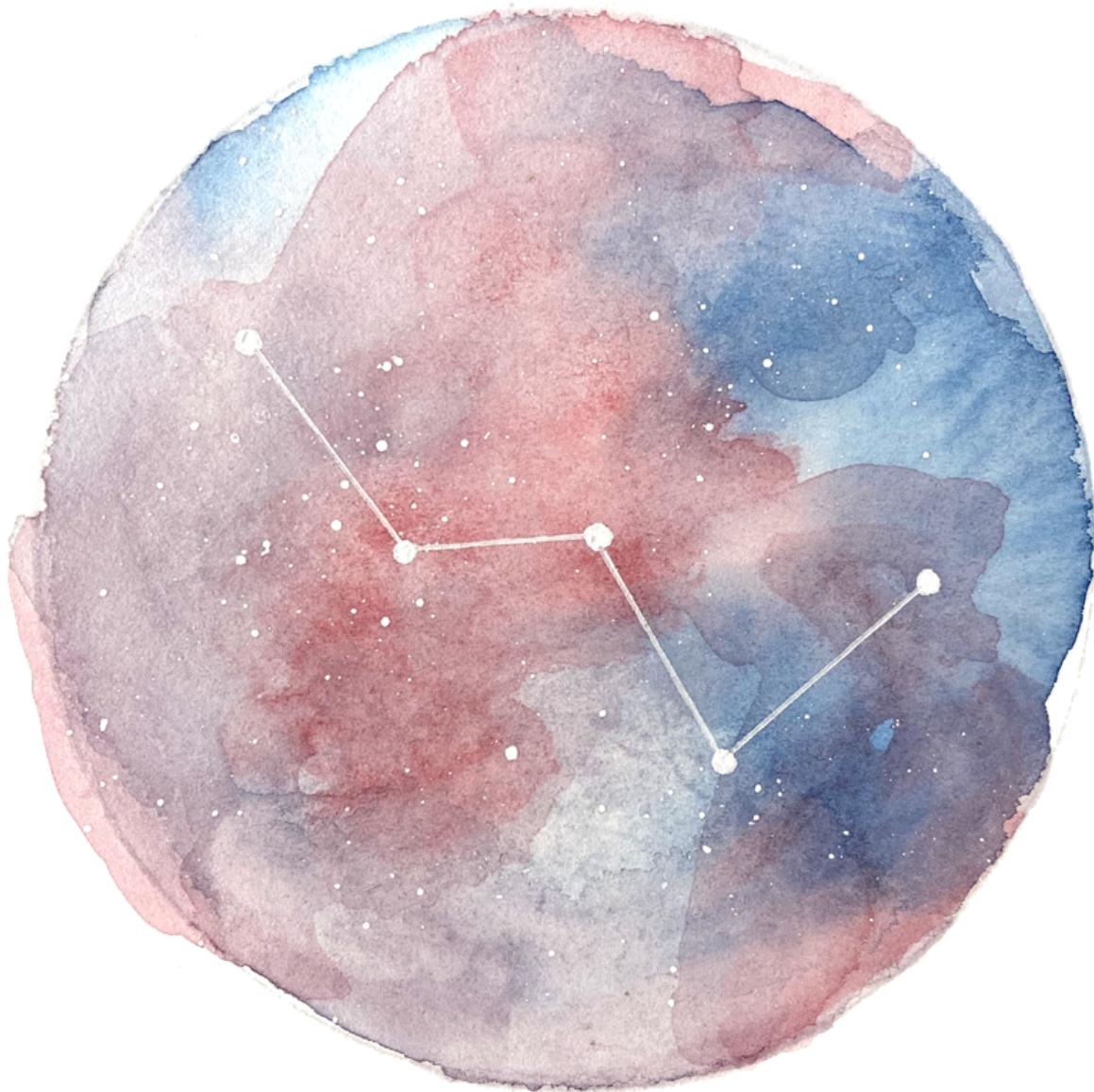
Inasmuch as I'm making a claim, you're right to want that sort of thing from me. But I'll disappoint you, for now, on this front. I cannot do very much at once; for me, just saying what I mean without misleading anyone is quite enough to be getting on with.

I will not try to argue that naturalism is important, either. Or, at least, not directly or on purpose. I won't say much of anything about when it matters, or why. This is also a worthwhile topic, but it's beyond the scope of this sequence.

Finally I will not try to help you learn naturalism. I *do* have a sometimes effective curriculum at this point, and I've even published [a sort of proto-naturalism introductory course](#) that you can take at your own pace online; but I will not be presenting anything like that here.

What I *will* try to do is pick out the concepts that are central to naturalism, name them with words, and tell you what I mean by those words.

It will take me seven-and-a-half essays, the first of which you have nearly finished.



When we are done here, I will write more things. When I write those things, I will sometimes use the term "naturalism". And if this sequence is successful, people who have read it will know what I'm talking about.

People who have not read this sequence will say "What is naturalism?", and I will finally be able to answer their question to my satisfaction.

Knowing the territory takes patient and direct observation. Let us begin, then, with “knowing”.

Alignment research exercises

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

It's currently hard to know where to start when trying to get better at thinking about alignment. So below I've listed a few dozen exercises which I expect to be helpful. They assume a level of background alignment knowledge roughly equivalent to what's covered in the [technical alignment track of the AGI safety fundamentals course](#). They vary greatly in difficulty - some are standard knowledge in ML, some are open research questions. I've given the exercises star ratings from * to *** for difficulty (note: *not* for length of time to complete - many require reading papers before engaging with them). However, I haven't tried to solve them all myself, so the star ratings may be significantly off.

I've erred on the side of including exercises which seem somewhat interesting and alignment-related even when I'm uncertain about their value; when working through them, you should keep the question "is this actually useful? Why or why not?" in mind as a meta-exercise. This post will likely be updated over time to remove less useful exercises and add new ones.

I'd appreciate any contributions of:

1. Comments about which exercises seem most or least useful.
2. Answers to the exercises.
3. More exercises! The ideal exercises are [nerdsnipe-style problems](#) which can be stated clearly, and seem well-defined, but lead into interesting depths when explored.

Reward learning

1. * Look at the examples of human feedback mechanisms discussed in the [reward-rational implicit choice paper](#). Think of another type of human feedback. What is the choice set? What is the grounding function?
 1. * [This paper by Anthropic](#) introduces a technique called context distillation. Describe this in terms of the reward-rational implicit choice framework.
 2. * Estimate the bandwidth of information conveyed by different types of human feedback. Describe a rough model for how this might change as training progresses. By contrast, how much information is conveyed by the choice of a programmatic reward function? (Consider both the case where the agent is given the exact reward function, and where it learns from reward observations.)
2. * Look at the examples of biases discussed in [learning the preferences of ignorant agents](#). Identify another bias which similarly influences human decision-making. Describe an example situation where a human with that bias might make the wrong decision. Formulate an algorithm that infers that human's true preferences.
 1. [Some answers here.](#)
3. ** Given that [humans can be assigned any values](#), why does reward learning ever work in practice?

4. ** Explain why [cooperative inverse reinforcement learning](#) doesn't solve the alignment problem.
 1. [Answer here.](#)

Agency

1. ** In [this paper](#), researchers devised a test for whether a recurrent network is doing planning: by seeing whether its performance improves when given more time to "think" before it can act. In the AlphaGo paper, researchers compared the performance of their MCTS+neural network algorithm against the network alone. Think of some other test that we could run that would give us evidence about the extent to which some neural network is internally doing planning.
2. * Consider [HCH](#), an attempted formalisation of "a human's enlightened judgment". Why might an implementation of HCH not be aligned? What assumptions would be needed to prevent that?
 1. *** In a later post, Paul defines a stronger version of HCH which "increases the complexity-theoretic expressiveness of HCH. The old version could be computed in EXPTIME, while the new version can compute any decidable function." Try to rederive a new version of HCH with these properties.
 2. [Answer here.](#)
3. * Ask the [OpenAI API](#) about what steps it would take to perform some long-term plan. Work in groups: think of a task that you expect it will be difficult to generate a good plan for, and then see who can design a prompt that will produce the best plan from the API.
 1. * Some steps of a plan generated by the API can also be performed by the API - e.g. a step which requires writing a poem about a given topic. What's the hardest task you can find for which the API can not only generate a plan, but also perform each of the steps in that plan?
4. ** [Pearl argues](#) that neural networks trained on supervised or self-supervised data can't learn to reason about interventions and counterfactuals ([see this post](#) for an explanation of the distinction). What's the strongest counterargument against his position?

Reinforcement learning

1. ** How is supervised learning on reward-maximising trajectories related (mathematically) to policy gradient with sparse, binary rewards?
2. ** What decision theories are implemented by different RL algorithms?
 1. [Some answers here.](#)
3. ** What might lead an RL agent to learn a policy which sacrifices reward in its current episode to get higher reward in a later episode?
 1. [Some answers in section 7 here.](#)
4. * Self-play in zero-sum two-player games converges to an optimal strategy (given sufficient assumptions about the model class). In other games, this isn't the case - why not?
5. ** Evaluate [this paper \(Reward is Enough\)](#). Does their argument hold up?
 1. ** After doing that: consider a bird practicing singing, which listens to its own song, and does RL using the rule: *the better the song sounds, the higher the reward*. But the bird is also deciding how much time to spend practicing singing versus foraging, etc. And the worse it sings, the more important it is to practice! So you really want the rule: *the worse the song*

sounds, the more rewarding it is to practice singing. How could you resolve this conflict?

2. [Some answers here.](#)

6. * Why can a behaviourally cloned policy perform well when run for a small set of timesteps, but poorly when run over a longer series of timesteps? How can this be fixed?
 1. [Some answers here.](#)
7. ** If a deep q-learning agent is trained in an environment where some actions lead to large negative rewards, it will never stop trying these actions (the policy will sometimes take these actions even when not randomly exploring due to epsilon exploration). Why does this happen? How could it be prevented?
 1. [Some answers here.](#)
8. ** RL agents have become capable of competent behaviour over longer and longer episodes. What difficulties arise in trying to measure improvements in how long they can act competently for? What metrics are most useful?
 1. The same question, but for sample efficiency rather than episode length.
 2. [Some answers here.](#)

Neural networks

1. * Consider [this paper on modularity in neural networks](#). Evaluate their metric of clustering; what others could we use instead?
2. ** Consider the following alignment proposal: a neural network has two output heads, one of which chooses actions, the other of which predicts the longer-term consequences of those actions. Suppose that we train the latter head to maximise human-evaluated prediction quality. What differences might we expect from backpropagating that loss all the way through the network, versus only backpropagating through the prediction head? What complications arise if we try to train the prediction head via RL? What advantages might there be of doing so?
3. ** “[Gradient hacking](#)” is a hypothesised phenomenon by which a model decides its actions partly on the basis of its observations of its own parameters, thereby changing the way its parameters are updated. Does the gradient hacking mechanism described in the linked post work? If not, does any variant of it work?
 1. [Some answers here.](#)
4. * Read [Jacob Steinhardt’s list](#) of examples of emergent shifts in machine learning. Can you think of any others? What about shifts that you expect in the near future?
5. ** What might it look like for the [circuits hypothesis](#) to be false?
6. * [This paper](#) discusses the metric of “effective data transferred”. What are the limitations of this metric? What are some alternative ways to measure data transfer?

Alignment theory

1. * Consider extending reinforcement learning to the case where rewards can depend on the parameters of a model. Why do classic convergence proofs no longer work?
 1. *** Are there any limiting assumptions which might lead to interesting theoretical results?

2. ** One concern with proposals to train using loss functions that depend directly on neural activations is that if we train a network to avoid carrying out any particular piece of cognition, that cognition will instead just be distributed across the network in a way that we can't detect. Describe a toy example of a cognitive trait that we can currently detect automatically. Design an experiment to determine whether, after training to remove that trait, the network has learned to implement an equivalent trait in a less-easily-detectable way.
3. *** Rederive some of the proofs in the following papers. For b) and c), explain what assumptions are being made about the optimality of the agents involved, and how they might break down in practice:
 1. [Seeking Power is Convergently Instrumental in MDPs](#)
 2. [AI safety via debate](#) (and see also [proofs about the effects](#) of adding [cross-examination](#))
 3. [Alignment proposals and complexity classes](#) (and this follow-up)
 4. [Some answers here](#).
4. *** Produce a proposal for [the ELK prize](#) (note that this requires engaging with the ELK writeup, which is very long).
5. ** Suppose that we're training a model via behavioural cloning of a human, but the human starts off with different prior knowledge to the model (either more knowledge, or less knowledge). How might this lead the model to behave in a misaligned way?
 1. [Some answers here](#).

Agent foundations

1. * [Open-source game theory](#)
2. ** [Selection theorems](#)
3. ** [Fixed point exercises](#)
4. *** [The 5 and 10 problem](#)

Evolution and economics

1. * An old study split insects into several groups which each lived together, and artificially selected in favour of smaller groups, in an attempt to study whether they would evolve to voluntarily restrain their breeding. Predict the outcome of the study.
 1. [Some answers here](#). Did the bias discussed in this post influence your expectations?
2. ** What might explain why there are so few hermaphroditic animal species, given that every individual being able to bear children could potentially double the number of children in the next generation?
 1. [Some answers here](#).
3. * Read this post about [evolving to extinction](#). Mathematically demonstrate that segregation-distorters could in fact lead a species to evolve to extinction.
4. * Evaluate [Fletcher and Doebeli](#)'s model of the evolution of altruism.
 1. Use the model to show how [the green-beard effect](#) could lead to the evolution of (a certain type of) altruism.
5. ** Why are roughly equal numbers of males and females born in most species?
 1. [Some answers here](#).
6. * Comparing GDP across time requires reference to [a standard basket of goods and services](#). What difficulties might this cause in taking GDP comparisons at

face value?

1. [Some answers here.](#)
7. ** Evaluate Roodman's model of explosive economic growth.
8. * In cooperative game theory, the "core" is the term for the set of allocations of payoffs to agents where no subset of the agents can form a coalition to improve their payoffs. For example, consider a group of N miners, who have discovered large bars of gold. Assume that two miners can carry one piece of gold, and so the payoff of any coalition S is $\text{floor}(|S|/2)$. If N is even, then the core consists of the single payoff distribution where each miner gets $\frac{1}{2}$. If N is odd, then the core is empty (because the miner who is left out can always make a better offer to some miner who currently has a gold-carrying partner). Identify the core for the following games:
 1. A game with 2001 players: 1000 of them have 1 left shoe, 1001 have 1 right shoe. A left-shoe/right-shoe pair can be sold for \$10.
 2. Mr A and Mr B each have three gloves. Any two gloves make a pair that they can sell for \$5.
 3. [Answers here.](#)
9. * How should coalitions decide how to split the payoffs they receive? The concept of [Shapley values](#) provides one answer. Convince yourself that Shapley values have the properties of linearity, null player and the stand-alone test described in the linked article.

Some important concepts in ML

These are intended less as exercises and more as pointers to open questions at the cutting edge of deep learning.

1. [Scaling laws](#)
 1. Why do they have the form they do?
 2. [Some answers here](#) and [here](#)
2. [Neural networks memorisation](#)
3. [Double descent](#)
4. [The lottery ticket hypothesis](#)
5. [Games with spinning-top structure](#)
6. [Gradient noise scale \(see also here\)](#)
7. [OpenAI Requests for Research](#)
8. [OpenAI Requests for Research 2](#)

Miscellaneous

1. * Fill in your estimates in [Cotra's timeline model](#). Does the model broadly make sense to you; are there ways you'd change it?
2. * Try playing [OpenAI's implementation of the Debate game](#).
3. ** Identify an important concept in alignment that isn't currently very well-explained; write a more accessible explanation.

Impossibility results for unbounded utilities

Some people think that they have unbounded utility functions. This isn't necessarily crazy, but it presents serious challenges to conventional decision theory. I think it probably leads to abandoning probability itself as a representation of uncertainty (or at least any hope of basing decision theory on such probabilities). This may seem like a drastic response, but we are talking about some pretty drastic inconsistencies.

This result is closely related to standard impossibility results in infinite ethics. I assume it has appeared in the philosophy literature, but I couldn't find it in the [SEP entry on the St. Petersburg paradox](#) so I'm posting it here. (Even if it's well known, I want something simple to link to.)

(ETA: this argument is extremely similar to Beckstead and Thomas' argument against Recklessness in [A paradox for tiny probabilities and enormous values](#). The main difference is that they use transitivity + "recklessness" to get a contradiction whereas I argue directly from "non-timidity." I also end up violating a dominance principle which seems even more surprising to violate, but at this point it's kind of like splitting hairs. I give a slightly stronger set of arguments in [Better impossibility results for unbounded utilities](#).)

Weak version

We'll think of preferences as relations $<$ over probability distributions over some implicit space of outcomes Ω (and we'll identify outcomes with the constant probability distribution). We'll show that there is no relation $<$ which satisfies three properties: Antisymmetry, Unbounded Utilities, and Dominance.

Note that we assume nothing about the existence of an underlying utility function. We don't even assume that the preference relation is complete or transitive.

The properties

Antisymmetry: It's never the case that both $A < B$ and $B < A$.

Unbounded Utilities: there is an infinite sequence of outcomes $X_1, X_2, X_4, X_8, \dots$ each "more than twice as good" as the last.^[1]

More formally, there exists an outcome X_0 such that:

- $X_{2k} < \frac{1}{2}X_{2k+1} + \frac{1}{2}X_0$ for every k .
- $X_0 < \frac{1}{2}X_1 + \frac{1}{2}X_0$ [2]

That is, X_1 is not as good as a $\frac{1}{2}$ chance of X_2 , which is not as good as a $\frac{1}{2}$ chance of X_4 , which is not as good as a $\frac{1}{2}$ chance of $X_8\dots$ This is nearly the weakest possible version of unbounded utilities.^[3]

Dominance: let A_0, A_1, \dots and B_0, B_1, \dots be sequences of lotteries, and p_0, p_1, \dots be a sequence of probabilities that sum to 1. If

$A_i < B_i$ for all i , then $\sum_i p_i A_i < \sum_i p_i B_i$.

Inconsistency proof

Consider the lottery

$$\begin{array}{ccccccc} X & & & = & & X & \\ & \infty & & & & 2 & \\ & & & - & & & \\ & & & & & 1 & \end{array}$$

We can write X_∞ as a mixture:

$$\bullet \quad X_\infty = \frac{1}{2}(X_0 + \frac{1}{2}X_1) + \frac{1}{4}(\frac{1}{2}X_0 + \frac{1}{2}X_2) + \frac{1}{8}(\frac{1}{2}X_0 + \frac{1}{2}X_4) + \dots$$

By definition $X_0 < \frac{1}{2}X_0 + \frac{1}{2}X_1$. And for each k , Unbounded Utilities implies that $X_{2k} < \frac{1}{2}X_0 + \frac{1}{2}X_{2k+1}$. Thus Dominance implies $X_\infty < X_\infty$, contradicting Antisymmetry.

How to avoid the paradox?

By far the easiest way out is to reject Unbounded Utilities. But that's just a statement about our preferences, so it's not clear we get to "reject" it.

Another common way out is to assume that any two "infinitely good" outcomes are incomparable, and therefore to reject Dominance.^[4] This results in being indifferent to receiving \$1 in every world (if the expectation is already infinite), or doubling the probability of all good worlds, which seems pretty unsatisfying.

Another option is to simply ignore small probabilities, which again leads to rejecting even the finite version of Dominance---sometimes when you mix together lotteries something will fall below the "ignore it" threshold leading the direction of your preference to reverse. I think this is pretty bizarre behavior, and in general ignoring small probabilities is much less appealing than rejecting Unbounded Utilities.

All of these options seem pretty bad to me. But in the next section, we'll show that if the unbounded utilities are symmetric---if there are both arbitrarily good *and* arbitrarily bad outcomes---then things get even worse.

Strong version

I expect this argument is also known in the literature; but I don't feel like people around LW usually grapple with exactly how bad it gets.

In this section we'll show there is no relation $<$ which satisfies three properties: Antisymmetry, Symmetric Unbounded Utilities, and Weak Dominance.

(ETA: actually I think that even with only positive utilities you already violate something very close to Weak Dominance, which [Beckstead and Thomas](#) call Prospect-Outcome dominance. I find this version of Weak Dominance slightly more compelling, but Symmetric Unbounded Utilities is a much stronger assumption than Unbounded Utilities or non-Timidity, so it's probably worth being aware of both versions. In a footnote ^[5] I also define an even weaker dominance principle that we are forced to violate.)

The properties

Antisymmetry: It's never the case that both $A > B$ and $B > A$.

Symmetric Unbounded Utilities. There is an infinite sequence of outcomes $X_1, X_{-2}, X_4, X_{-8}, \dots$ each of which is "more than twice as important" as the last but with opposite sign. More formally, there is an outcome X_0 such that:

- $X_0 < X_1$
- For every even k : $\frac{1}{2}X_{-2^{k+1}} + \frac{1}{2}X_{2^k} < X_0$
- For every odd k : $\frac{1}{2}X_{2^{k+1}} + \frac{1}{2}X_{-2^k} > X_0$

That is, a certainty of X_1 is outweighed by a $\frac{1}{2}$ chance of X_{-2} , which is outweighed by a $\frac{1}{2}$ chance of X_4 , which is outweighed by a $\frac{1}{2}$ chance of X_{-8}

Weak Dominance.^[5] For any outcome X , any sequence of lotteries B_0, B_1, \dots , and any sequence of probabilities p_0, p_1, \dots that sum to 1:

- If $X < B_i$ for every i , then $X < \sum p_i B_i$.
- If $X > B_i$ for every i , then $X > \sum p_i B_i$.

Inconsistency proof

Now consider the lottery $X_{\pm\infty} = \frac{1}{2}X_1 + \frac{1}{2}X_{-2} + \frac{1}{4}X_4 + \frac{1}{16}X_{-8} + \frac{1}{32}X_{16} + \dots$

We can write $X_{\pm\infty}$ as the mixture:

$$\bullet \frac{1}{2}(\frac{1}{2}X_1 + \frac{1}{2}X_{-2}) + \frac{1}{4}(\frac{1}{2}X_4 + \frac{1}{2}X_{-8}) + \frac{1}{8}(\frac{1}{2}X_{16} + \frac{1}{2}X_{-32}) \dots$$

By Unbounded Utilities each of these terms is $< X_0$. So by Weak Dominance, $X_{\pm\infty} < X_0$.

But we can also write $X_{\pm\infty}$ as the mixture:

- $\frac{1}{2} X_1 + \frac{1}{2} (\frac{1}{2} X_{-2} + \frac{1}{2} X_4) + \dots (\frac{1}{2} X_{-8} + \frac{1}{2} X_{16}) + \dots$

By Unbounded Utilities each of these terms is $> X_0$. So by Weak Dominance $X_{\pm\infty} > X_0$. This contradicts Antisymmetry.

Now what?

As usual, the easiest way out is to abandon Unbounded Utilities. But if that's just the way you feel about extreme outcomes, then you're in a sticky situation.

You could allow for unbounded utilities as long as they only go in one direction. For example, you might be open to the possibility of arbitrarily bad outcomes but not the possibility of arbitrarily good outcomes.^[6] But the asymmetric version of unbounded utilities doesn't seem very intuitively appealing, and you *still* have to give up the ability to compare any two infinitely good outcomes (violating Dominance).

People like talking about extensions of the real numbers, but those don't help you avoid any of the contradictions above. For example, if you want to extend $<$ to a preference order over hyperreal lotteries, it's just even *harder* for it to be consistent.

Giving up on Weak Dominance seems pretty drastic. At that point you are *talking* about probability distributions, but I don't think you're really using them for decision theory--it's hard to think of a more fundamental axiom to violate. Other than Antisymmetry, which is your other option.

At this point I think the most appealing option, for someone committed to unbounded utilities, is actually much more drastic: I think you should give up on probabilities as an abstraction for describing uncertainty, and should not try to have a preference relation over lotteries at all.^[7] There are no ontologically fundamental lotteries to decide between, so this isn't necessarily so bad. Instead you can go back to talking directly about preferences over uncertain states of affairs, and build a totally different kind of machinery to understand or analyze those preferences.

ETA: replacing dominance

Since writing the above I've become more sympathetic to violations of Dominance and even Weak Dominance---it would be pretty jarring to give up on them, but I can at least imagine it. I still think violating "Very Weak Dominance"^[8] is pretty bad, but I don't think it captures the full weirdness of the situation.

So in this section I'll try to replace Weak Dominance by a principle I find even more robust: if I am indifferent between X and *any* of the lotteries A_i , then I'm also indifferent between X and any mixture of the lotteries A_i . This isn't strictly weaker than Weak Dominance, but violating it feels even weirder to me. At any rate, it's another fairly strong impossibility result constraining unbounded utilities.

The properties

We'll work with a relation \leq over lotteries. We write $A = B$ if both $A \leq B$ and $B \leq A$. We write $A < B$ if $A \leq B$ but not $A = B$. We'll show that \leq can't satisfy four properties: Transitivity, Intermediate mixtures, Continuous symmetric unbounded utilities, and Indifference to homogeneous mixtures.

Intermediate mixtures. If $A < B$, then $A < \frac{1}{2}A + \frac{1}{2}B < B$.

Transitivity. If $A \leq B$ and $B \leq C$ then $A \leq C$.

Continuous symmetric unbounded utilities. There is an infinite sequence of lotteries $X_1, X_{-2}, X_4, X_{-8}, \dots$ each of which is "exactly twice as important" as the last but with opposite sign. More formally, there is an outcome X_0 such that:

- $X_0 < X_1$
- For every even k : $\frac{1}{2}X_{-2^{k+1}} + \frac{1}{2}X_{2^k} = X_0$
- For every odd k : $\frac{1}{2}X_{2^{k+1}} + \frac{1}{2}X_{-2^k} = X_0$

That is, a certainty of X_1 is exactly offset by a $\frac{1}{2}$ chance of X_{-2} , which is exactly offset by a $\frac{1}{4}$ chance of X_4 , which is exactly offset by a $\frac{1}{8}$ chance of X_{-8}

Intuitively, this principle is kind of like symmetric unbounded utilities, but we assume that it's possible to dial down each of the outcomes in the sequence (perhaps by mixing it with X_0) until the inequalities become exact equalities.

Homogeneous mixtures. Let X be an outcome, A_0, A_1, A_2, \dots , a sequence of lotteries, and p_0, p_1, p_2, \dots be a sequence of probabilities summing to 1. If $A_i = X$ for all i , then $\sum p_i A_i = X$.

Inconsistency proof

Consider the lottery $X_{\pm\infty} = \frac{1}{2}X_1 + \frac{1}{4}X_{-2} + \frac{1}{8}X_4 + \frac{1}{16}X_{-8} + \frac{1}{32}X_{16} + \dots$

We can write $X_{\pm\infty}$ as the mixture:

$$\bullet \frac{1}{2}(\frac{1}{2}X_1 + \frac{1}{2}X_{-2}) + \frac{1}{16}(\frac{1}{2}X_4 + \frac{1}{2}X_{-8}) + \frac{1}{64}(\frac{1}{2}X_{16} + \frac{1}{2}X_{-32}) \dots$$

By Unbounded Utilities each of these terms is $= X_0$. So by homogeneous mixtures, $X_{\pm\infty} = X_0$.

But we can also write $X_{\pm\infty}$ as the mixture:

$$\bullet \frac{1}{2}X_1 + \frac{1}{8}(\frac{1}{2}X_{-2} + \frac{1}{2}X_4) + \frac{1}{32}(\frac{1}{2}X_{-8} + \frac{1}{2}X_{16}) + \dots$$

By Unbounded Utilities each of these terms other than the first is $= X_0$. So by Homogenous Mixtures, the combination of all terms other than the first is $= X_0$. Together with the fact that $X_1 > X_0$, Intermediate Mixtures and Transitivity imply $X_{\pm\infty} > X_0$. But that contradicts $X_{\pm\infty} = X_0$.

1. [^](#)

Note that we could replace "more than twice as good" with "at least 0.00001% better" and obtain exactly the same result. You may find this modified version of the principle more appealing, and it is closer to non-timidity as defined in [Beckstead and Thomas](#). Note that the modified principle implies the original by applying transitivity 100000 times, but you don't actually need to apply transitivity to get a contradiction, you can just apply Dominance to a different mixture.

2. [^](#)

You may wonder why we don't just write $X_0 < X_1$. If we did this, we'd need to introduce an additional assumption that if $A < B$, $pA + (1 - p)X_0 < pB + (1 - p)X_0$. This would be fine, but it seemed nicer to save some symbols and make a slightly weaker assumption.

3. [^](#)

The only plausibly-weaker definition I see is to say that there are outcomes $X_0 < X_1$ and an infinite sequence $X_{>2}, X_{>3}, \dots$ such that for all n : $\frac{1}{2}X_{>n} + (1 - \frac{1}{2})X_0 > X_1$. If we replaced the $>$ with $=$ then this would be stronger than our version, but with the inequality it's not actually sufficient for a paradox.

To see this, consider a universe with three outcomes X_0, X_1, X_∞ and a preference order $<$ that always prefers lotteries with higher probability of X_∞ and breaks ties using by preferring a higher probability of X_1 . This satisfies all of our other properties. It satisfies the weaker version of the axiom by taking $X_{>n} = X_\infty$ for all n , and it wouldn't be crazy to say that it has "unbounded" utilities.

4. [^](#)

For realistic agents who think unbounded utilities are possible, it seems like they should assign positive probability to encountering a St. Petersburg paradox such that [all decisions have infinite expected utility](#). So this is quite a drastic thing to give up on. See also: [Pascal's mugging](#).

5. [^](#)

I find this principle pretty solid, but it's worth noting that the same inconsistency proof would work for the even weaker "Very Weak Dominance": for any pair of outcomes with $X_0 < X_1$, and any sequence of lotteries B_i each strictly better than X_1 , any mixture of the B_i should at least be strictly better than X_0 !

6. [^](#)

Technically you can also violate Symmetric Unbalanced Utility while having both arbitrarily good *and* arbitrarily bad outcomes, as long as those outcomes aren't comparable to one another. For example, suppose that worlds have a real-valued amount of suffering and a real-valued amount of pleasure. Then we could have a lexical preference for minimizing expected suffering (considering all worlds with infinite expected suffering as incomparable), and try to maximize pleasure only as a tie-breaker (considering all worlds with infinite expected pleasure as incomparable).

7.

Instead you could keep probabilities but abandon infinite probability distributions. But at this point I'm not exactly sure what unbounded utilities means---if each decision involves only finitely many outcomes, then in what sense do all the other outcomes exist? Perhaps I may face infinitely many possible decisions, but each involves only finitely many outcomes? But then what am I to make of my parent's decisions while raising me, which affected my behavior in each of those infinitely many possible decisions? It seems like they face an infinite mixture of possible outcomes. Overall, it seems to me like giving up on infinitely big probability distributions implies giving up on the spirit of unbounded utilities, or else going down an even stranger road.

Learning By Writing

I have very detailed opinions on lots of topics. I sometimes get asked how I do this, which might just be people making fun of me, but I choose to interpret it as a real question, and I'm going to sketch an answer here.

You can think of this as a sort of sequel to [Minimal-Trust Investigations](#). That piece talked about how investigating things in depth can be valuable; this piece will try to give a sense of how to get an in-depth investigation off the ground, going from "I've never heard of this topic before" to "Let me tell you all my thoughts on that."

The rough basic idea is that I organize my learning around **writing** rather than reading. This doesn't mean I don't read - just that the reading is always in service of the writing.

Here's an outline:

| | |
|--------|--|
| Step 1 | Pick a topic |
| Step 2 | Read and/or discuss with others (a bit) |
| Step 3 | Explain and defend my current, incredibly premature hypothesis, in writing (or conversation) |
| Step 4 | Find and list weaknesses in my case |
| Step 5 | Pick a subquestion and do more reading/discussing |
| Step 6 | Revise my claim / switch sides |
| Step 7 | Repeat steps 3-6 a bunch |
| Step 8 | Get feedback on a draft from others, and use this to keep repeating steps 3-6 |

The "traditionally" hard parts of this process are steps 4 and 6: spotting weaknesses in arguments, trying to resist the temptation to "stick to my guns" when my original hypothesis isn't looking so good, etc.

But step 3 is a different kind of challenge: trying to "always have a hypothesis" and re-articulating it whenever it changes. By doing this, I try to **continually focus my reading on the goal of forming a bottom-line view, rather than just "gathering information."** I think this makes my investigations more focused and directed, and the results easier to retain. I consider this approach to be **probably the single biggest difference-maker between "reading a ton about lots of things, but retaining little" and "efficiently developing a set of views on key topics and retaining the reasoning behind them."**

Below I'll give more detail on each step, then some brief notes (to be expanded on later) on why this process is challenging.

My process for learning by writing

Step 1: pick a topic. First, I decide what I want to form an opinion about. My basic approach here is: "Find claims that are important if true, and might be true."

This doesn't take creativity. We live in an ocean of takes, pundits, advocates, etc. I usually cheat by paying special attention to claims by people who seem particularly smart, interesting, unconventionally minded (not repeating the same stuff I hear everywhere), and interested in the things I'm interested in (such as the [long-run future of humanity](#)).

But I also tend to be at least *curious* about *any* claim that is both "important if true" and "not obviously wrong according to some concrete reason I can voice," even if it's coming from a very random source (Youtube commenter, whatever).

For a concrete example throughout this piece, I'll use this hypothesis, which I examined pretty recently: "Human history is a story of life getting gradually, consistently better."

(Other, more complicated examples are the [Collapsing Civilizational Competence Hypothesis](#); the [Most Important Century hypothesis](#); and my [attempt to summarize history in one table](#).)

Step 2: read and/or discuss (a bit). I usually start by trying to read the most prominent 1-3 pieces that (a) defend the claim or (b) attack the claim or (c) set out to comprehensively review the evidence on both sides. I try to understand the major reasons they're giving for the side they come down on. I also chat about the topic with people who know more about it than I do, and who aren't too high-stakes to chat with.

In the example I'm using, I read the relevant parts of [Better Angels of our Nature](#) and [Enlightenment Now](#) (focusing on claims about life getting better, and skipping discussion of "why"). I then looked for critiques of the books that specifically responded to the claims about life having gotten better (again putting aside the "why"). This led mostly to [claims about the peacefulness of hunter-gatherers](#).

Step 3: explain and defend my current, incredibly premature hypothesis, in writing (or conversation). This is where my approach gets unusual - I form a hypothesis about whether the claim is true, LONG before I'm "qualified to have an opinion." The process looks less like "Read and digest everything out there on the topic" and more like "Read the 1-3 most prominent pieces on each side, then go."

I don't have an easy time explaining "how" I generate a hypothesis while knowing so little - it feels like I just always have a "guess" at the answer to some topic, whether or not I even want to (though it often takes me a lot of effort to *articulate* the guess in words). The main thing I have to say about the "how" is that it just **doesn't matter**: at this stage the hypothesis is more about setting the stage for more questions about investigation than about really trying to be right, so it seems sufficient to "just start rambling onto the page, and make any corrections/edits that my current state of knowledge already forces."

For this example, I noted down something along the lines of: "Life has gotten better throughout history. The best data on this comes from the last few hundred years, because before that we just didn't keep many records. Sometimes people try to claim that the longest-ago, murkiest times were better, such as [hunter-gatherer times](#), but there's no evidence for this - in fact, empirical evidence shows that hunter-gatherers were very violent - and we should assume that these early times fit on the same general trendline, which would mean they were quite bad. (Also, if you go even further back than hunter-gatherers, you get to apes, whose lives seem really horrible, so that seems to fit the trend as well.¹)"

It took real effort to disentangle the thoughts in my head to the point where I could write that, but I tried to focus on keeping things simple and not trying to get it perfect.

At this stage, this is **not** a nuanced, caveated, detailed or well-researched take. Instead, my approach is more like: “Try to state what I think in a pretty strong, bold manner; defend it aggressively; list all of the best counterarguments, and shoot them down.” **This generally fails almost immediately.**

Step 4: find and list weaknesses in my case. My next step is to play devil’s advocate against myself, such as by:

- Looking for people arguing things that contradict my working hypothesis, and looking for their strongest points.
- Noting claims I’ve made with this property: “I haven’t really made an attempt to look comprehensively at the arguments on both sides of this, and if I did I might change my mind.”

(This summary obscures an ocean of variation. Having more existing knowledge about a general area, and more experience with investigations in general, can make someone much better at noticing things like this.)

In the example, my “devil’s advocate” points included:

- I’m getting all of my “life has gotten better” charts from books that are potentially biased. I should do something to see whether there are other charts, excluded from those books, that tell the opposite story.
- From my brief skim, the “hunter-gatherers were violent” claim looks right, and the critiques seem very hand-wavy and non-data-based. But I should probably read them more carefully and pull out their strongest arguments.
- Even if hunter-gatherers were violent, what about other aspects of their lives? [Wikipedia](#) seems to have a pretty rosy picture ...

In theory, I could swap Step 4 (listing things I’d like to look into more) with Step 3 (writing what I think). That is, I could try to review both sides of every point comprehensively before forming my own view, which means a lot more reading before I start writing.

I think many people try to do this, but in my experience at least, it’s not the best way to go.

- Debates tend to be many-dimensional: for example, “Has life gotten better?” quickly breaks down into “Has quality-of-life metric X gotten better over period Y?” for a whole bunch of different X-Y pairs (plus other questions²).
- So if my goal were “Understand both sides of every possible sub-debate,” I could be reading forever - for example, I might get embroiled in the debates and nuances around each different claim made about life getting better over the last few hundred years.
- By writing early, I get a chance to make sure I’ve written down the *version of the claim I care most about*, and make sure that any further investigation is focused on the things that matter most for changing my mind on this claim.
 - Once I wrote down “There are a huge number of charts showing that life has gotten better over the last few hundred years,” I could see that deep-diving any particular one of those charts wouldn’t be the best use of time - compared to addressing the very weakest points in the claim I had written,

by going back further in time to hunter-gatherer periods, or looking for entirely different collections of charts.

Step 5: pick a subquestion and do more reading and/or discussing. One of the most important factors that determines whether these investigations go well (in the sense of teaching me a lot relatively quickly) is **deciding which subquestions to “dig into” and which not to**. As just noted, writing the hypothesis down early is key.

I try to stay very focused on doing the reading (and/or low-stakes discussion) most likely to change the big-picture claim I’m making. I rarely read a book or paper “once from start to finish”; instead I energetically skip around trying to find the parts most likely to give me a solid reason to change my mind, read them carefully and often multiple times, try to figure out what else I should be reading (whether this is “other parts of the same document” or “academic papers on topic X”) to contextualize them, etc.

Step 6: Revise my claim / switch sides. This is one of the trickiest parts - pausing Step 5 as soon as I have a modified (often still simplified, under-researched and wrong) hypothesis. It’s hard to notice when my hypothesis changes, and hard to stay open to radical changes of direction (and I make no claim that I’m as good at it as I could be).

I often try radically flipping around my hypothesis, even if I haven’t actually been convinced that it’s wrong - sometimes when I’m feeling iffy about arguing for one side, it’s productive to just go ahead and try arguing for the other side. **I tend to get further by noticing how I feel about the “best arguments for both sides” than by trying from the start to be even-handed.**

In the example, I pretty quickly decided to try flipping my view around completely, and noted something like: “A lot of people assume life has gotten better over time, but that’s just the last few hundred years. In fact, our best guess is that hunter-gatherers were getting some really important things right, such as gender relations and mental health, that we still haven’t caught up to after centuries of progress. Agriculture killed that, and we’ve been slowly climbing out of a hole ever since. There should be tons more research on what hunter-gatherer societies are/were like, and whether we can replicate their key properties at scale today - this is a lot more promising than just continuing to push forward science and technology and modernity.”

This completely contradicted my initial hypothesis. (I [now think both are wrong](#).)

This sent me down a new line of research: constructing the best argument I could that life was better in hunter-gatherer times.

Step 7: repeat steps 3-6 a bunch. I tried to gather the best evidence for hunter-gatherer life being good, and for it being bad, and zeroed in on gender relations and violence as particularly interesting, confusing debates; on both of these, I changed my hypothesis/headline several times.

My hypotheses became increasingly complex and detailed, as you can see from the final products: [Pre-agriculture gender relations seem bad](#) (which argues that gender relations for hunter-gatherers were/are far from Wikipedia’s rosy picture, according to the best available evidence, though the evidence is far from conclusive, and it’s especially unclear how pre-agriculture gender relations compare to today’s) and [Unraveling the evidence about violence among very early humans](#) (which argues that

hunter-gatherer violence was indeed high, but that - contra *Better Angels* - it probably got even worse after the development of agriculture, before declining at some pretty unknown point before today).

I went through several cycles of “I think I know what I really think and I’m ready to write,” followed by “No, having started writing, I’m unsatisfied with my answer on this point and think a bit more investigation could change it.” So I kept alternating between writing and reading, but was always reading with the aim of getting back to writing.

I finally produced some full, opinionated drafts that seemed to me to be about the best I could do without a ton more work.

After I had satisfied myself on these points, I popped back up from the “hunter-gatherer” question to the original question of whether life has gotten better over time. I followed a similar process for investigating other subquestions, like “Is the set of charts I’ve found representative for the last few hundred years?” and “What about the period in between hunter-gatherer times and the last few hundred years?”

Step 8: add feedback from others into the loop. It takes me a long time to get to the point where I can no longer easily tear apart my own hypothesis. Once I do, I start seeking feedback from others - first just people I know who are likely to be helpful and interested in the topic, then experts and the public. This works the same basic way as Steps 4-7, but with others doing a lot of the “noticing weaknesses” part (Step 4).

When I publish, I am thinking of it more like “I can’t easily find more problems with this, so it’s time to see whether others can” than like “This is great and definitely right.”

I hope I haven’t made this sound fun or easy

Some things about this process that are hard, taxing, exhausting and a bit of a mental health gauntlet:

- I constantly have a feeling (after reading) like I know what I think and how to say it, then I start writing and immediately notice that I don’t at all. I need to take a lot of breaks and try a lot of times to even “write what I currently think,” even when it’s pretty simple and early.
- Every subquestion is something I could spend a lifetime learning about, if I chose to. I need to constantly interrupt myself and ask, “Is this a key point? Is this worth learning more about?” or else I’ll never finish.
- There are infinite tough judgment calls about things like “whether to look into some important-seeming point, or just reframe my hypothesis such that I don’t need to.” Sometimes the latter is the answer (it feels like some debate is important, but if I really think about it, I realize the thing I most care about can be argued for without getting to the bottom of it); sometimes the former is (it feels like I can try to get around some debate, but actually, I can’t really come to a reasonable conclusion without an exhausting deep dive).
- At any given point, I know that if I were just better at things like “noticing which points are really crucial” and “reformulating my hypothesis so that it’s easier to defend while still important,” I could probably do something twice as good in half the time ... and I often realize after a massive deep dive that most of the time I spent wasn’t necessary.

- Because of these points, I have very little ability to predict when a project will be done; I am never confident that I'm doing it as well as I could; and I'm constantly interrupting myself to reflect on these things rather than getting into a flow.
- Half the time, all of this work just ends up with me agreeing with conventional wisdom or “the experts” anyway ... so I’ve just poured in work and gone through a million iterations of changing my mind, and any random person I talk to about it will just be like “So you decided X? Yeah X is just what I had already assumed.”
- The whole experience is a mix of writing, Googling, reading, skimming, and pressuring myself to be more efficient, which is very different and much more unpleasant compared to the experience of just reading. (Among other things, I can read in a nice location and be looking at a book or e-ink instead of a screen. Most of the work of an “investigation” is in front of a glowing screen and requires an Internet connection.)

I'll write more about these challenges in a future post. I definitely recommend reading as a superior leisure activity, but for me at least, writing-centric work seems better for learning.

I'm really interested in comments from anyone who tries this sort of thing out and has things to share about how it goes!

Footnotes

1. I never ended up using this argument about apes. I think it's probably mostly right, but there's a whole can of worms with claims about loving, peaceful bonobos that I never quite got motivated to get to the bottom of. [↩](#)
2. Such as which metrics are most important. [↩](#)

Theses on Sleep

Published originally (with all of the footnotes) on my site: <https://guzey.com/theses-on-sleep/>

Summary: In this essay, I question some of the consensus beliefs about sleep, such as the need for at least 7 hours of sleep for adults, harmfulness of acute sleep deprivation, and harmfulness of long-term sleep deprivation and our inability to adapt to it.

It appears that the evidence for all of these beliefs is much weaker than sleep scientists and public health experts want us to believe. In particular, I conclude that it's plausible that at least acute sleep deprivation is not only not harmful but beneficial in some contexts and that it's that we are able to adapt to long-term sleep deprivation.

I also discuss the bidirectional relationship of sleep and mania/depression and the costs of unnecessary sleep, noting that sleeping 1.5 hours per day less results in gaining more than a month of wakefulness per year, every year.

*Note: I sleep the normal 7-9 hours if I don't restrict my sleep. However, stimulants like coffee, modafinil, and adderall seem to have much smaller effect on my cognition than on cognition of most people I know. My brain in general, as you might guess from reading [my site](#), is not very normal. So, be cautious before trying anything with your sleep on the basis of the arguments I lay out below. Specifically **do not** make any drastic changes to your sleep schedule on the basis of reading this essay and, if you want to experiment with sleep, do it gradually (i.e. varying the average amount of sleep by no more than 30 minutes at a time) and carefully.*

Comfortable modern sleep is an unnatural superstimulus. Sleepiness, just like hunger, is normal.

The default argument for sleeping 7-9 hours a night is that this is the amount of sleep most of us get "naturally" when we sleep without using alarms. In this section, I argue against this line of reasoning, using the following analogy:

1. Experiencing hunger is normal and does not necessarily imply that you are not eating enough. Never being hungry means you are probably eating too much.
 2. Experiencing sleepiness is normal and does not necessarily imply that you are undersleeping. Never being sleepy means you are probably sleeping too much.
-

Most of us (myself included) eat a lot of junk food and candy if we don't restrict ourselves. Does this mean that lots of junk food and candy is the "natural" or the "optimal" amount for health?

Obviously, no. Modern junk food and candy are unnatural superstimuli, much tastier and much more abundant than any natural food, so they end up overwhelming our brains with pleasure, especially given that we are bored at work, college, or in high school so much of the day.

What if the only food available to you was junk food and candy?

1. If you don't eat any, you starve.

2. If you eat just enough to be lean, you'll keep salivating at the sight of pizzas and ice cream and feel distracted and hungry all the time. Importantly, in this situation, the feeling of hunger does not mean that you should eat more - it's your brain being overpowered by a superstimulus while being bored.
3. If you eat it as much as you want, you'll probably eat too much and become fat.
 - And if you eat way too much candy or pizza at once, you'll be feeling terrible afterwards, however tasty the food was.

Most of us (myself included) sleep 7-9 hours if we don't have any alarms in the morning and if we get out of bed when we feel like it. Does this mean that 7-9 hours of sleep is the "natural" or the "optimal" amount?

My thesis is: obviously, no. Modern sleep, in its infinite comfort, is an unnatural superstimulus that overwhelms our brains with pleasure and comfort (note: I'm not saying that it's bad, simply that being in bed today is much more pleasurable than being in "bed" in the past.)

Think about sleep 10,000 years ago. You sleep in a cave, in a hut, or under the sky, with predators and enemy tribes roaming around. You are on a wooden floor, on an animal's skin, or on the ground. The temperature will probably drop 5-10°C overnight, meaning that if you were comfortable when you were falling asleep, you are going to be freezing when you wake up. Finally, there's moon shining right at you and all kinds of sounds coming from the forest around you.

In contrast, today: you sleep on your super-comfortable machine-crafted foam of the exact right firmness for you. You are completely safe in your home, protected by thick walls and doors. Your room's temperature stays roughly constant, ensuring that you stay warm and comfy throughout the night. Finally, you are in a light and sound-insulated environment of your house. And if there's any kind of disturbance you have eye masks and earplugs.

Does this sound "natural"?

Now, what if the only sleep available to you was modern sleep?

1. If you don't sleep at all, you go crazy, because some amount of sleep is necessary.
2. If you sleep just enough to be awake during the day, you'll be dreaming of getting a nap at the sight of a bed and will be distracted and sleepy all the time. Importantly, I claim, in this situation, the feeling of sleepiness does not mean that you should sleep more - it's your brain being overpowered by a superstimulus while being bored.
3. **I claim that if you sleep as much as you want, you'll probably sleep too much and become more susceptible to depression.**
 - And if you sleep way too much at once, you'll be feeling terrible afterwards, however pleasant the sleep was.

Even if I convinced you about the "sleeping too much" part, you are still probably wondering: but what does depression have to do with anything? Isn't sleeping a lot good for mental health? Well...

Depression <-> oversleeping. Mania <-> acute sleep deprivation

In this section, I argue that depression triggers/amplifies oversleeping while oversleeping triggers/amplifies depression. Similarly, mania triggers/amplifies acute sleep deprivation while acute sleep deprivation triggers/amplifies mania.

One of the most notable facts about sleep is just how interlinked excessive sleep is with depression and how interlinked sleep deprivation is with mania in bipolar people.

Someone in r/BipolarReddit asked: [How many hours do you sleep when stable vs \(hypo\)manic? Depressed?](#)

Here are all 8 answers that compare hours for manic and depressed states, note the consistency:

- “Manic/hypomanic: 0-6 hours Stable: 7-9 hours Depressed: 10-19 hours”
- “Manic, 2-3, hypo, 5-6, stable 8-9, depressed 10-12. 8 is the number I try to hit.”
- “Severely depressed w/o mixed features - 12 to 15 hours
Low to Moderate depressed w/o mixed - 10 hours, if no alarm. With alarm less, but super hangover
Stable -Usually 7-9 hours
Hypomanic taking sedating evening meds - 5 to 7 hours
Hypomanic with no sedating evening meds - 3 to 5 hours
Manic out of hand - 0 to 3 hours
Manic in hospital put on maximum sedating meds or injections - 4 to 6 hours
Mixed episodes = same as hypo(manic)”
- “I try to get at least 8 hours but when I’m depressed I nap a lot. When I’m hypo I sleep pretty much the same but when I’m manic I’m lucky to get 3 hours. Huhs”
- “Just got out of a manic episode. A few all-nighters, a lot of 3 hour nights, and a good night of sleep was 6 hours. Now I’m depressed and I’ve been sleeping from 9pm to noon and staying in bed for much longer after I’m awake.”
- “Manic 2-4, stable 6-7, depressed 10-12”
- “Around 15 hours of sleep per night while depressed, and between 0-4 hours per night while manic.”

Lack of sleep is such a potent trigger for mania that [*acute sleep deprivation is literally used to treat depression.*](#) Aside from ketamine, not sleeping for a night is the only medicine we have to quickly – literally overnight – and reliably (in ~50% of patients) improve mood in depressed patients (until they go to bed, unless you [*keep advancing*](#) their sleep phase).
NOTE: DO NOT TRY THIS IF YOU ARE BIPOLAR, YOU MIGHT GET A MANIC EPISODE.

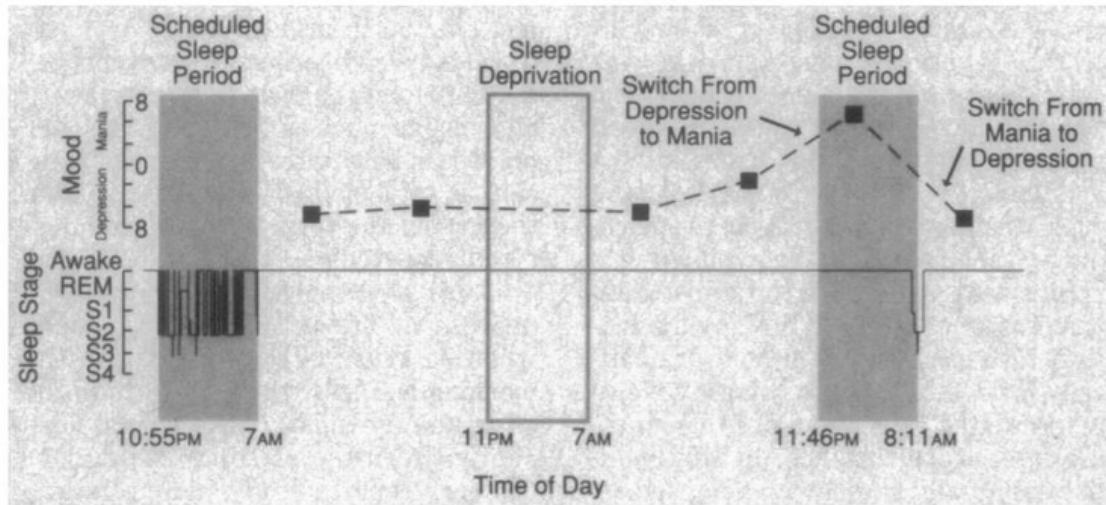


Fig 2.—Precipitation of mania following one night of total sleep deprivation and relapse into depression following recovery sleep in a depressed patient with bipolar illness. Sleep was recorded polygraphically and scored according to conventional sleep-stage criteria. Despite having been awake for more than 40 hours, the patient had difficulty returning to sleep because she had switched into a manic state. After 43 minutes of mostly stage 2 sleep (right), she awoke depressed.

Figure 1. Copied from Wehr TA. Improvement of depression and triggering of mania by sleep deprivation. JAMA. 1992 Jan 22;267(4):548-51.

Why does the lack of sleep promote manic states while long sleep promotes depression? I don't know. But here are a couple of pointers to interesting papers relevant to the question: [Can non-REM sleep be depressogenic?](#) Brain-derived neurotrophic factor (BDNF) is associated with synapse growth. Sleep deprivation appears to increase BDNF [and therefore neurogenesis?]. Papers that showed up when I googled "sleep deprivation bdnf": [The Brain-Derived Neurotrophic Factor: Missing Link Between Sleep Deprivation, Insomnia, and Depression. The link between sleep, stress and BDNF. BDNF: an indicator of insomnia?](#) [Recovery Sleep Significantly Decreases BDNF In Major Depression Following Therapeutic Sleep Deprivation.](#)

[Jeremy Hadfield](#) writes:

My (summarized/simplified) hypothesis based on what I've read: depression involves rigid, non-flexible brain states that correspond to rigid depressive world models. Depression also involves a non-updating of models or inability to draw new connections (brain is even literally slightly lighter in depressed patients). Sleep involves revising/simplifying world models based on connections learned during the day, involves pruning unneeded or irrelevant synaptic connections. Thus, excessive sleep + depression = even less world model updating, even more rigid brain, even fewer new connections. Sleep deprivation can resolve this problem at least temporarily by ensuring that you stay awake for longer and keep adding connections, thus compensating for the decreased connection-building caused by depression and "forcing" a brain update (perhaps through neural annealing - see QRI article).

Occasional acute sleep deprivation is good for health and promotes more efficient sleep

In this section, I continue the analogy between eating and sleeping and extend it to exercise. I ask: if fasting and exercising are good, shouldn't acute sleep deprivation also be good? And I conclude that it *is* probably good.

Let's continue our analogy of sleep to eating and add exercise to the mix.

It seems to me that most common arguments against acute sleep deprivation equally "demonstrate" that fasting and exercise are bad.

For example, I ran 7 kilometers 2 days ago and my legs still hurt like hell and I can't run at all. Does this mean that running is "bad"?

Well, consensus seems to be that dizziness, muscle damage (and thus pain) and decreased physical performance after the run, are not just not bad, but are in fact necessary for the organism to train to run faster or to run longer distances by increasing muscle mass, muscle efficiency, and lung capacity.

What about fasting? When I fast, I am more anxious, I think about food a lot, meaning that focus is more difficult, and I feel cold. And if I decided to fast too much, I would pass out and then die. Does this mean that fasting is "bad"? Well, consensus seems to be that occasional fasting actually activates some "good" kind of stress, promotes healthy autophagy, (obviously) helps to lose weight, etc. and is in fact *good*.

Now, what happens when I sleep for 2 hours instead of 7 one night? I feel somewhat tingly in my hands, my mood is heightened a little bit, and, if I start watching a movie with my wife at 6pm, I'll fall asleep. Does this mean that sleeping 2 hours one night is bad for my health?

Obviously no. **The only thing we observe is that my organism was subjected to acute stress.** However, the reaction to acute stress does not tell us anything about the long-term effects of this kind of stress. As we know, both in running and in fasting, short-term acute stress response results in adaptation and in long-term increase in performance and in benefit to the organism.

I combed through a *lot* of sleep literature and I haven't seen a single study that made a parallel to either fasting or exercise and I haven't seen a single pre-registered RCT that tried to see what happens to someone if you subject them to 1-3 nights per week of acute sleep deprivation and allow to recover the rest of the nights. Do they perform better or worse in the long-term on cognitive tests? Do they have more or less inflammation? Do they need less recovery sleep over time?

I think that the answers are:

1. Acute sleep deprivation combined with caffeine or some other stimulant that cancels out sleep pressure does not result in decreased cognitive ability at least until 30-40 hours of wakefulness (if this is true, then *sleepiness*, rather *absence of sleep* per se is responsible for decreased cognitive performance during acute sleep deprivation).
2. Occasional acute sleep deprivation has no impact on long-term cognitive ability or health.
3. Sleep *does* become more efficient over time and, in complete analogy to exercise, you withstand both acute sleep deprivation better and can function at baseline with a lower amount of sleep in the long-term.

(The only parallel to fasting I'm aware of anyone making is by Nassim Taleb... when he was [quote-tweeting](#) me.)

Also see:

- [Appendix: anecdotes about acute sleep deprivation](#)
- [Appendix: Philipp Streicher on homeostasis, its relationship to mania/depression, and on other points I make](#)

Our priors about sleep research should be weak

In this section, I note that most sleep research is extremely unreliable and we shouldn't conclude much on the basis of it.

[Do you believe in power-posing? In ego depletion? In hungry judges and brain training?](#)

If the answer is no, then your priors for our knowledge about sleep should be weak because "sleep science" is mostly just rebranded cognitive psychology, with the vast majority of it being small-n, not pre-registered, p-hacked experiments.

I have been able to find exactly *one* pre-registered experiment of the impact of prolonged sleep deprivation on cognition. It was published by economists from Harvard and MIT in 2021 and its pre-registered analysis found [null or negative effects of sleep on all primary outcomes](#) (note that both the abstract and the main body of this paper report results without the multiple-hypothesis correction, in contradiction to the pre-registration plan of the study. The paper does not mention this change anywhere.).

So why has sleep research not been facing a severe replication crisis, similar to psychology?

First, compared to psychology, where you just have people fill out questionnaires, sleep research is slow, relatively expensive, and requires specialized equipment (e.g. EEG, actigraphs). So skeptical outsiders go for easier targets (like social psychology) while the insiders keep doing the same shoddy experiments because they need to keep their careers going *somewhere*.

Second, imagine if sleep researchers had conclusively shown that sleep is not important for memory, health, etc. – would they get any funding? No. Their jobs are literally predicated on convincing the NIH and other grantmakers that sleep is important. As Patrick McKenzie [notes](#), "If you want a problem solved make it someone's project. If you want it managed make it someone's job."

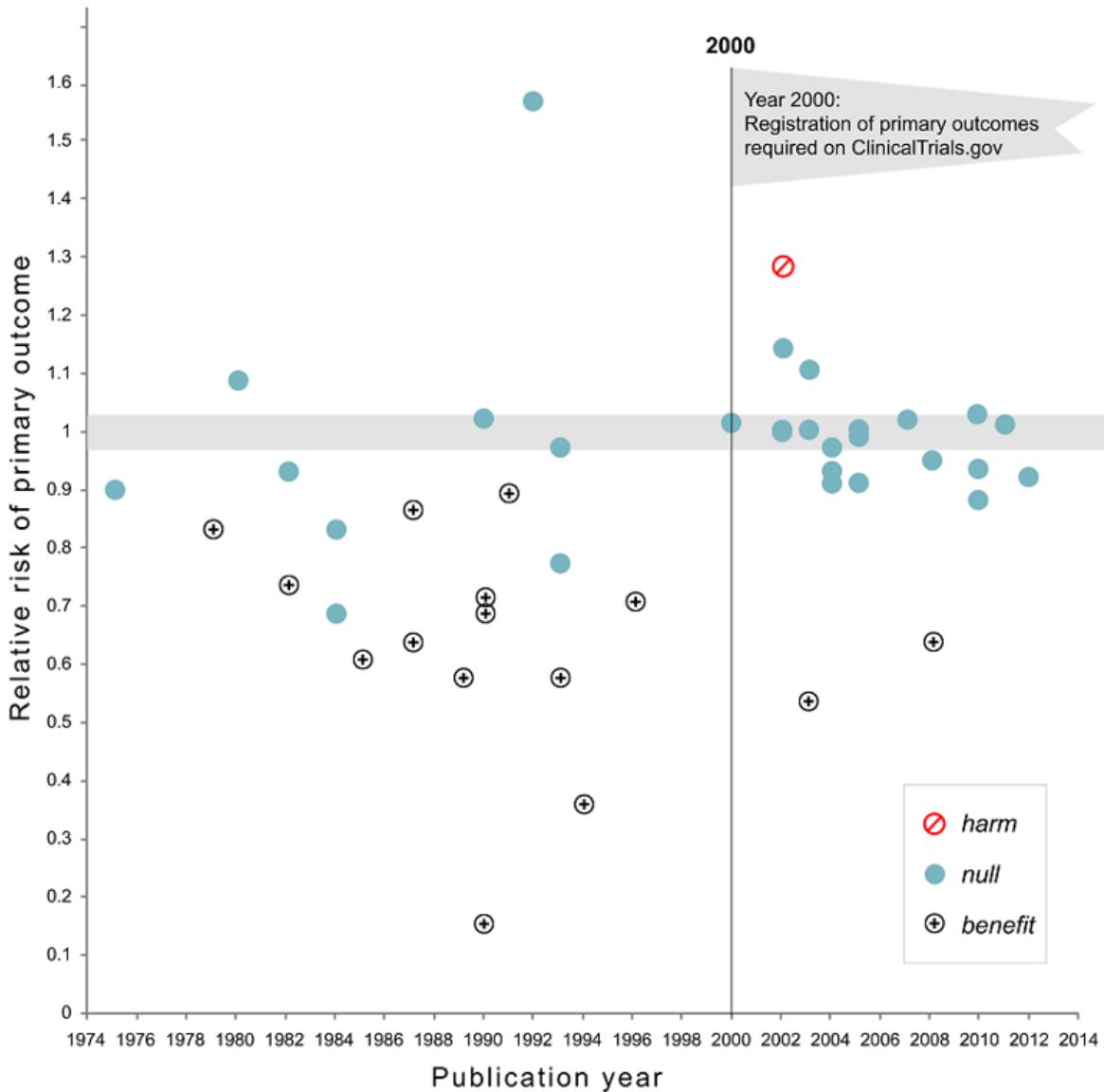


Figure 2. Relative risk of showing benefit or harm of treatment by year of publication for large NHLBI trials on pharmaceutical and dietary supplement interventions. Copied from Kaplan RM, Irvin VL. Likelihood of null effects of large NHLBI clinical trials has increased over time. PloS one. 2015 Aug 5;10(8):e0132382.

Even in medicine, without pre-registered RCTs truth is extremely difficult to come by, with [more than one half](#) of high-impact cancer papers failing to be replicated, and with one half of RCTs without pre-registration of positive outcomes [being spun](#) by researchers as providing benefit when there's none. And this is in *medicine*, which is infinitely more consequential and rigorous than psychology.

Also see: [Appendix: I have no trust in sleep scientists.](#)

Decreasing sleep by 1-2 hours a night in the long-term has no negative health effects

In this section, I outline several lines of evidence that bring me to the conclusion that decreasing sleep by 1-2 hours a night in the long-term has no negative health effects. To summarize:

1. A sleep researcher who trains sailors to sleep efficiently in order to maximize their race performance believes that 4.5-5.5 hours of sleep is fine.
 2. 70% of 84 hunter-gatherers studied in 2013 slept less than 7 hours per day, with 46% sleeping less than 6 hours.
 3. A single-point mutation can decrease the amount of required sleep by 2 hours, with no negative side-effects.
 4. A brain surgery can decrease the amount of sleep required by 3 hours, with no negative-side effects.
 5. Sleep is not required for memory consolidation.
-

1. Claudio Stampi is a Newton, Massachusetts based sleep researcher. But he is not your normal sleep researcher whose career is built on observational studies or p-hacked n=20 experiments that always show “significant” results. He is one of the only sleep researchers with skin in the game: the goal of his research is to maximize performance of sailors by tinkering with their sleep cycles, and he believes that 4.5-5.5 hours of sleep is fine, as long as it’s broken down into core sleep and a series of short (usually 20-minute) naps. Here’s [Outside](#):

“Solo sailing is one of the best models of 24/7 activity, and brains and muscles are required,” Stampi said one day at his home, from which he runs the institute. “If you sleep too much, you don’t win. If you don’t sleep enough, you break.” ...

“For those sailors who are seriously competing, Stampi is a necessity,” says Brad Van Liew, a 37-year-old Californian who began working with Stampi in 1998 and went on to become America’s most accomplished solo racer and the winner in his class of the 2002-2003 Around Alone, a 28,000-mile global solo race. “You have to sleep efficiently, or it’s like having a bad set of sails or a boat bottom that isn’t prepared properly.” ...

both Golding and MacArthur sleep about the same amount while racing, between 4.5 and 5.5 hours on average in every 24—the minimum amount, Stampi believes, on which humans can get by.

In 2013, scientists [tracked](#) the sleep of 84 hunter-gatherers from 3 different tribes (each person’s sleep was measured for about a week but measurements for different groups were taken in different parts of the year). **The average amount of sleep among these 84 people was 6.5 hours.** Judging by CDC’s “7 hours or more” [recommendation](#), 70% out of these 84 undersleep:

- 6 people slept between 4 and 5 hours
- 19 people slept between 5 and 6 hours
- 34 people slept between 6 and 7 hours
- 21 people slept between 7 and 8 hours
- 4 people slept between 8 and 9 hours

It appears that there is a distinct single-point mutation that allows some people to sleep several hours less than typical on average. [A Rare Mutation of β1-Adrenergic Receptor Affects Sleep/Wake Behaviors](#):

We have identified a mutation in the β1-adrenergic receptor gene in humans who require fewer hours of sleep than most. In vitro, this mutation leads to decreased protein stability and dampened signaling in response to agonist treatment. In vivo, the mice carrying the same mutation demonstrated short sleep behavior. We found that this receptor is highly expressed in the dorsal pons and that these ADRB1+ neurons are active during rapid eye movement (REM) sleep and wakefulness. Activating these

neurons can lead to wakefulness, and the activity of these neurons is affected by the mutation. These results highlight the important role of β 1-adrenergic receptors in sleep/wake regulation.

The study compares carriers of the mutation in one family to non-carriers in the same family and finds that carriers sleep about 2 hours per day less. Given the complexity of sleep and the multitude of its functions, it seems extremely implausible that just one mutation in the β 1-adrenergic receptor gene was able to increase its efficiency by about 25%. It seems that it just made carriers sleep less (due to more stimulation of a group of neurons in the brain responsible for sleep/wakefulness) without anything else obviously changing when compared to non-carriers.

A similar example of a drop in the amount of sleep required without negative side effects and driven by a single factor was described in [Development of a Short Sleeper Phenotype after Third Ventriculostomy in a Patient with Ependymal Cysts](#). To sum up: a 59-year-old patient had chronic hydrocephalus. An endoscopic third ventriculostomy was performed on him. His sleep dropped from 7-8 hours a night to 4-5 hours a night without him becoming sleepy, he stopped being depressed, and his physical or cognitive performance stayed normal, as measured by the doctors.

Sleep is not required for memory consolidation. Jerome Siegel (the author of the hunter-gatherers study mentioned above) writes in [Memory Consolidation Is Similar in Waking and Sleep:](#)

Under interference conditions, such as exist during sleep deprivation, subjects, by staying awake, necessarily interacting with the experimenter keeping them awake and experiencing the laboratory environment, will remember more than just the items that are presented. But they may be less able to recall the particular items the experimenter is measuring. This can lead to the mistaken conclusion that sleep is required for memory consolidation [7].

Recent work has, for the first time, dealt with this issue. It was shown that a quiet waking period or a meditative waking state in which the environment is being ignored, produces a gain in recall similar to that seen in sleep, relative to an active waking state or a sleep-deprived state [8-16]. ...

REM sleep has been hypothesized to have a key role in memory consolidation [20]. But it has been reported that near total REM sleep deprivation for a period of 14 to 40 days by administration of the monoamine oxidase inhibitor phenelzine has no apparent effect on cognitive function in humans [21]. **A systematic study using serotonin or norepinephrine re-uptake inhibitors to suppress REM sleep in humans had no deleterious effects on a variety of learning tasks [22, 23].** Humans rarely survive the damage to the pontine region which when discretely lesioned in animals greatly reduces or eliminates REM sleep [20, 23-25]. However, one such subject with pontine damage that severely reduced REM sleep has been thoroughly studied. The studies show normal or above normal cognitive performance and no deficit in memory formation or recall [26•]. **It has been claimed that learning results in greater total amounts of sleep, or greater amounts of REM sleep [27], or greater amounts of sleep spindles, or slow wave activity. However, a systematic test of this hypothesis in 929 human subjects with night-long EEG recording found no such correlation with retention [28•].**

The entire Scientific Consensus™ about sleep being essential for memory consolidation appears to be heavily flawed, driven by [people like Matthew Walker, and making me lose the last remnants of trust in sleep science that I had.](#)

Also see:

- [Appendix: how I wake up after 6 or less hours of sleep](#)
- [Appendix: anecdotes about long-term sleep deprivation](#)
- [Appendix: the idea that sleep's purpose is metabolite clearance, if not total bs, is massively overhyped](#)

Conclusion

Chadwick worked for several nights straight without sleep on the seminal discovery [of the neutron, for which he was awarded the 1935 Nobel in physics]. When he was done he went to a meeting of the Kapitza Club at Cambridge and gave a talk about it, ending with the words, “Now I wish to be chloroformed and put to sleep”.

[Ash Jogalekar](#)

I'm not what they call a "natural short sleeper". If I don't restrict my sleep, I often sleep more than 8 hours and I still struggle with getting out of bed. I used to be *really* scared of not sleeping enough and almost never set the alarm for less than 7.5 hours after going to bed.

My sleep statistics tells me that I slept an average of 5:25 hours over the last 7 days, 5:49 hours over the last 30 days, and 5:57 over the last 180 days hours, meaning that I'm awake for 18 hours per day instead of 16.5 hours. I usually sleep 5.5-6 hours during the night and take a nap a few times a week when sleepy during the day.

This means that **I'm gaining 33 days of life every year.** 1 more year of life every 11 years. 5 more years of life every 55 years.

Why are people not all over this? Why is everyone in love with [charlatans](#) who say that sleeping 5 hours a night will double your risk of cancer, make you pre-diabetic, and cause Alzheimer's, despite studies showing that people who sleep 5 hours have the same, if not lower, mortality than those who sleep 8 hours? Convincing a million 20-year-olds to sleep an unnecessary hour a day is equivalent, in terms of their hours of wakefulness, to killing 62,500 of them.

I wrote large chunks of this essay having slept less than 1.5 hours over a period of 38 hours. I came up with and developed the biggest arguments of it when I slept an average of 5 hours 39 minutes per day over the preceding 14 days. At this point, I'm pretty sure that the entire "not sleeping 'enough' makes you stupid" is a 100% psyop . It makes you somewhat more sleepy, yes. More stupid, no. I literally did an experiment in which I tried to find changes in my cognitive ability [after sleeping 4 hours a day for 12-14 days](#), I couldn't find any. My friends who I was talking to a lot during the experiment simply didn't notice anything.

What do I lose due to sleeping 1.5 hours a day less? I'm somewhat more sleepy every day and staying awake during boring calls is even more difficult now. There's no guarantee that what I'm doing is healthy after all, although, as I explained above, I think that it's extremely unlikely due to likely adaptation, and likely beneficial effects of sleep deprivation (e.g. increased BDNF, less susceptibility to depression), and since I take a 20-minute nap under my wife's watch whenever I don't feel good.



ALIEN SOLDIER

@_alien_soldier

...

Replying to @milken_cookies and @alexeyguzey

It's a known scientific fact there is zero loss of cognitive function for a 7AM start after spending an entire night out so long as you down Red Bull every hour and take a shower before heading in.

Too many snowflakes and not enough fun these days.

10:15 PM · May 15, 2021 · Twitter for iPhone

9 Likes

An internationally known expert on acute sleep deprivation Dr. ALIEN SOLDIER (twitter account deleted).

Acknowledgements

I would like to thank (in reverse alphabetic order): [Misha Yagudin](#), Bart Sturm, [Ulysse Sabbag](#), [Gavin Leech](#), [Stephen Malina](#), [Anastasia Kuptsova](#), [Jake Koenig](#), Aleksandr Kotyurgin, [Alexander Kim](#), [Basil Halperin](#), [Jeremy Hadfield](#), [Steve Gadd](#), and [Willy Chertman](#) for reading drafts of this essay and for disagreeing with many parts of it vehemently. All errors mine.

Citation

Cite as:

Guzey, Alexey. Theses on Sleep. Guzey.com. 2022 February. Available from <https://guzey.com/theses-on-sleep/>.

Or [download a BibTeX file here](#).

Notes

- One popular sleep tip I've come to wholeheartedly believe is the importance of waking up at the same time: from my experience, it does really seem that the organism adjusts the time it is ready to wake up if you keep a consistent schedule.
- I think sleepiness indicates boringness of the environment much more than it indicates the physiological need for sleep. It's an indicator of build up of sleep-promoting chemicals coupled with the boringness of the environment

- observation: I find staying awake during boring lectures impossible and reliably fall asleep during them, regardless of the amount of sleep I'm getting
- observation: I can play video games with little sleep for several days and feel 100% alert (a superstimulus of its own, but still a valuable observation)
- observation: I become sleepy when I'm working on something boring and difficult

Common objections

Objection: "When I'm underslept I notice that I'm less productive."

Answer: It might be that undersleeping itself causes you to be less productive. However, it might also be the case that there's an upstream cause that results in both undersleeping and lack of productivity. I think either could be the case depending on the person but understanding what exactly happens is much harder than people typically appreciate when they notice such co-occurrence's.

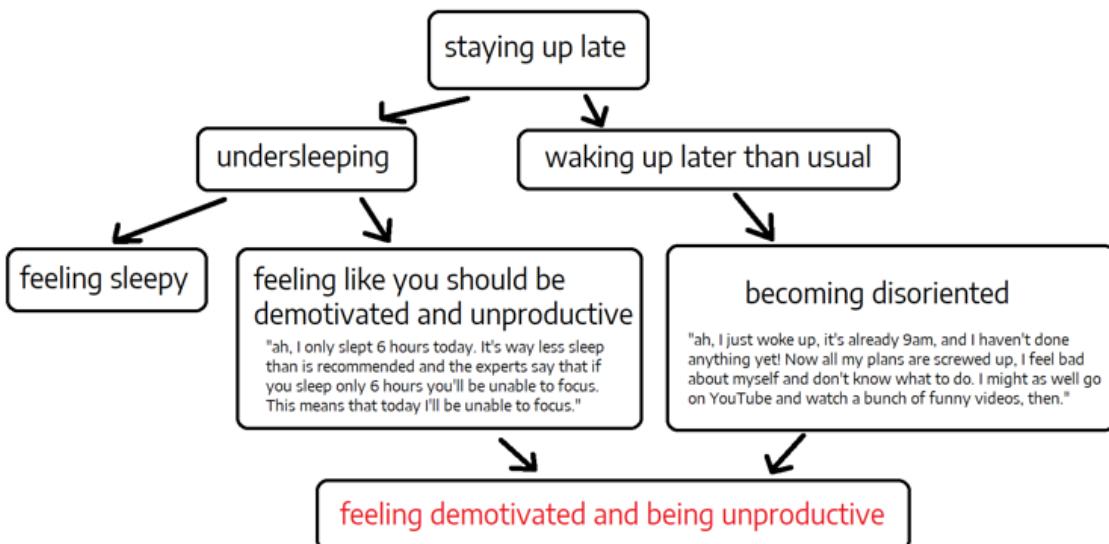


Figure 5. Causal graph of the "staying up late and feeling demotivated and being unproductive scenario."

Objection: "Driving when you are sleepy is dangerous, therefore you are wrong."

Answer: Yep, I agree that driving while being sleepy is dangerous and I don't want anyone to drive, to operate heavy machinery, etc. when they are sleepy. This, however, bears no relationship on any of the arguments I make.

Objection: "The graph that shows more sleep being associated with higher doesn't tell us anything because sick people tend to sleep more."

Answer: It is true that some diseases lead to prolonged sleep. However, some diseases also lead to shortened sleep. For example, [many stroke patients suffer from insomnia](#) and people with fatal familial insomnia struggle with insomnia. Therefore, if you want to make the argument that the association between longer sleep and higher mortality is not indicative of the effect of sleep, you have to accept that the same is true about shorter sleep and higher mortality.

Appendix: I have no trust in sleep scientists

Why do I bother with all of this theorizing? Why do I think I can discover something about sleep that thousands of them couldn't discover over many decades?

The reason is that I have approximately 0 trust in the integrity of the field of sleep science.

As you might be aware, 2 years ago I wrote [a detailed criticism of the book Why We Sleep](#) written by a Professor of Neuroscience at psychology at UC Berkeley, the world's leading sleep researcher and the most famous expert on sleep, and the founder and director of the Center for Human Sleep Science at UC Berkeley, Matthew Walker.

Here are just a few of biggest issues (there were many more) with the book.

Walker wrote: "Routinely sleeping less than six or seven hours a night demolishes your immune system, more than doubling your risk of cancer", despite there being no evidence that cancer in general and sleep are related. There are obviously no RCTs on this, and, in fact, [there's not even a correlation between general cancer risk and sleep duration.](#)

[Walker falsified a graph from an academic study in the book.](#)

Walker outright fakes data to support his "sleep epidemic" argument. The data on sleep duration Walker presents on the graph below simply [does not exist:](#)

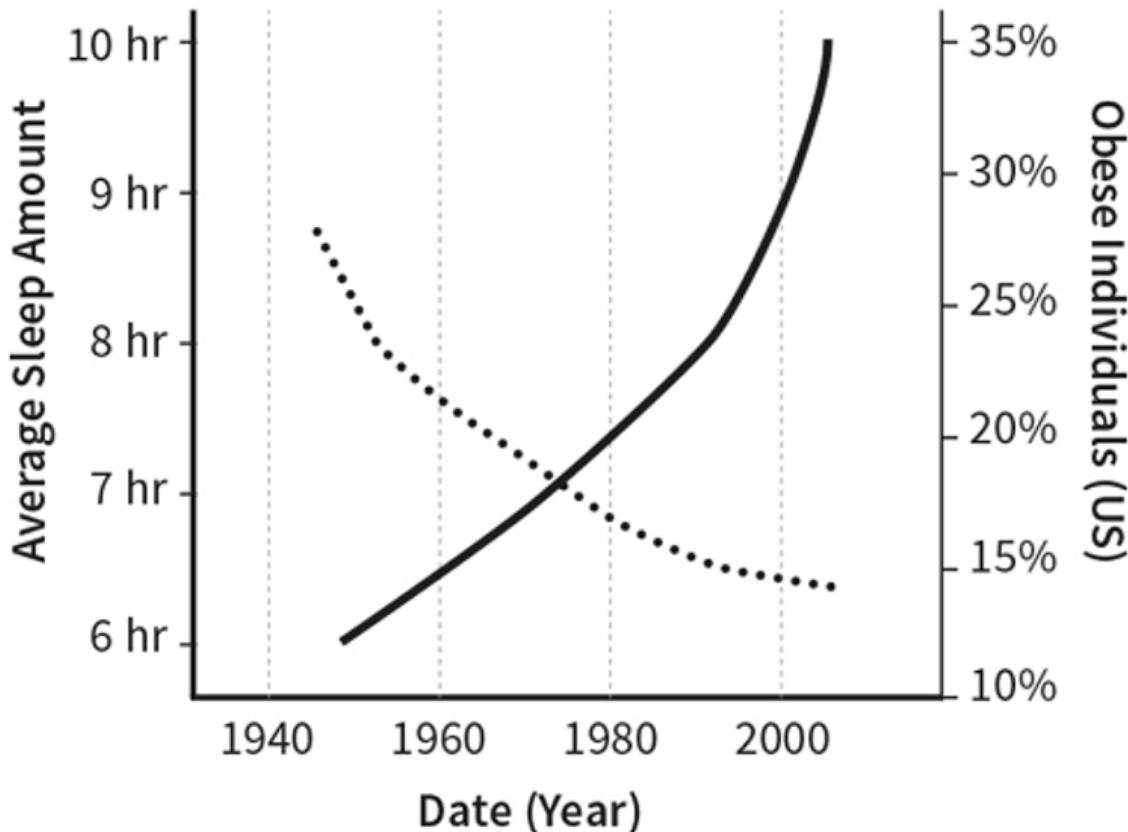


Figure 6. Sleep loss and obesity. Country not specified for sleep data. Copied from Walker M. Why we sleep: Unlocking the power of sleep and dreams. Simon and Schuster; 2017 Oct 3.

Here's some actual data on sleep duration over time:

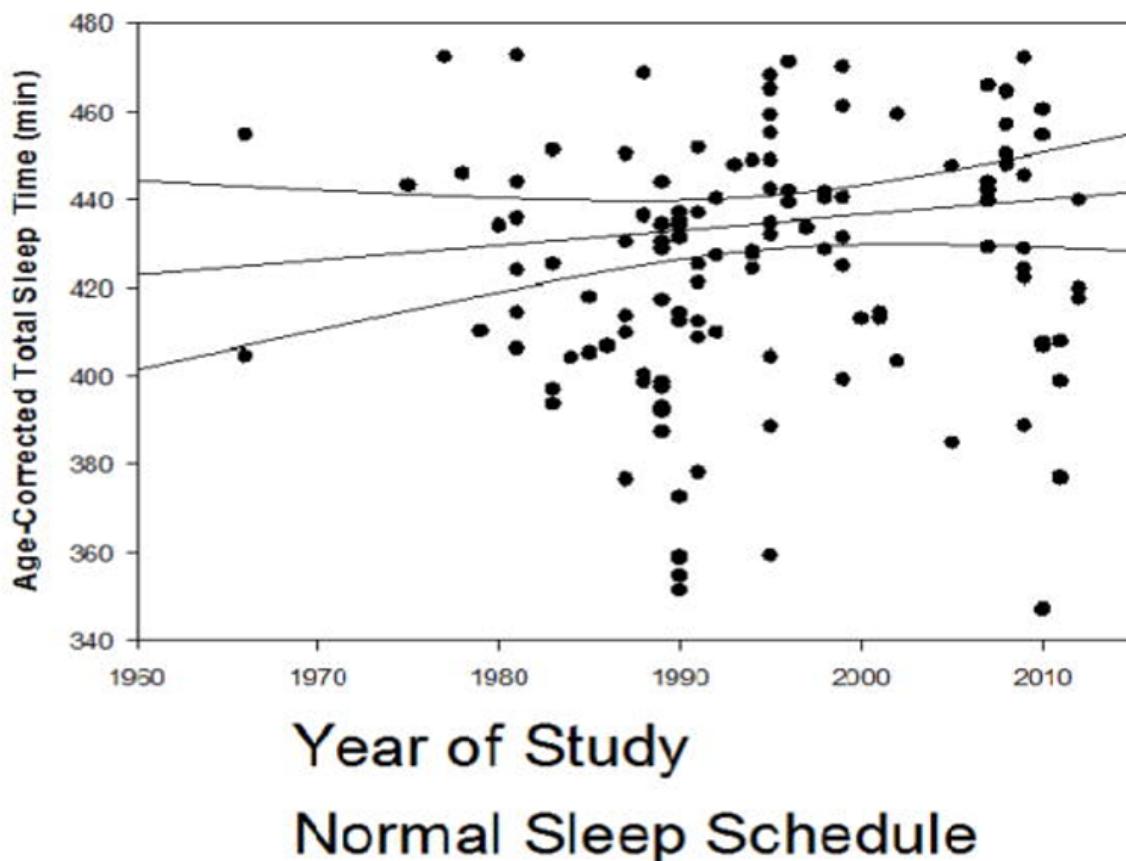


Figure 7. Association of year of study with age-adjusted total sleep time (min) for studies in which subjects followed their usual sleep schedule. Copied from Youngstedt SD, Goff EE, Reynolds AM, Kripke DF, Irwin MR, Bootzin RR, Khan N, Jean-Louis G. Has adult sleep duration declined over the last 50+ years?. Sleep medicine reviews. 2016 Aug 1;28:69-85.

By the time my review was published, the book had sold hundreds of thousands if not millions of copies and was praised by the [New York Times](#), [The Guardian](#), and many other highly-respected papers. It was [named one of NPR's favorite books of 2017](#) while Walker went on a full-blown podcast tour.

Did any sleep scientists voice the concerns they with the book or with Walker? No. They were too busy listening to his keynote at the Cognitive Neuroscience Society 2019 meeting.

Did any sleep scientists voice their concerns after I published my essay detailing its errors and fabrications? No (unless you count people replying to me on Twitter as “voicing a concern”).

Did Walker lose his status in the community, his NIH grants, or any of his appointments? No, no, and no.

I don't believe that a community of scientists that refuses to police fraud and of which Walker is a foremost representative could be a community of scientists that would produce a trustworthy and dependable body of scientific work.

Appendix: the idea that sleep's purpose is metabolite clearance, if not total bs, is massively overhyped

Specifically, [the original 2013 paper](#) accumulated more than 3,000 (!) citations in less than 10 years and is highly misleading.

The paper is called "Sleep Drives Metabolite Clearance from the Adult Brain". The abstract says:

The conservation of sleep across all animal species suggests that sleep serves a vital function. We here report that sleep has a critical function in ensuring metabolic homeostasis. Using real-time assessments of tetramethylammonium diffusion and two-photon imaging in live mice, we show that natural sleep or anesthesia are associated with a 60% increase in the interstitial space, resulting in a striking increase in convective exchange of cerebrospinal fluid with interstitial fluid. In turn, convective fluxes of interstitial fluid increased the rate of β -amyloid clearance during sleep. Thus, the restorative function of sleep may be a consequence of the enhanced removal of potentially neurotoxic waste products that accumulate in the awake central nervous system.

At the same time, the paper found that anesthesia *without sleep* results in the same clearance (paper: "A β clearance did not differ between sleeping and anesthetized mice"), meaning that clearance is not caused by sleep per se, but instead only co-occurs with it. Authors did not mention this in the abstract and mistitled the paper, thus misleading the readers. As far as I can tell, literally nobody pointed this out previously.

And on top of all of this "125I-A β 1-40 was injected intracortically", meaning that they did not actually find any brain waste products that would be cleared out. This is an exogenous compound that was injected god knows where disrupting god knows what in the brain.

Appendix: anecdotes about acute sleep deprivation

Max Levchin in *Founders at Work*:

The product wasn't really finished, and about a week before the beaming at Buck's I realized that we weren't going to be able to do it, because the code wasn't done. Obviously it was really simple to mock it up—to sort of go, "Beep! Money is received." But I was so disgusted with the idea. We have this security company; how could I possibly use a mock-up for something worth \$4.5 million? What if it crashes? What if it shows something? I'll have to go and commit ritual suicide to avoid any sort of embarrassment. So instead of just getting the mock-up done and getting reasonable rest, my two coders and I coded nonstop for 5 days. I think some people slept; I know I didn't sleep at all. It was just this insane marathon where we were like, "We have to get this thing working." It actually wound up working perfectly. The beaming was at 10:00 a.m.; we were done at 9:00 a.m.

[/u/CPlusPlusDeveloper](#) on Gwern's [Writing in the Morning](#):

We know that acute sleep deprivation seems to have a manic and euphoric effect on at least some percent of the population some percent of the time. For example staying up all night is one of the most effective ways to temporarily alleviate depression. Of course the

problem is that chronic sleep deprivation has the opposite effect, and the temporary mania and euphoria is not sustainable.

My speculative take is that whatever this mechanism, it was the main reason you experienced a productivity boost. By waking up early you intentionally were fighting against your chronobiology, hence adding an element of acute sleep deprivation regardless of how many hours you got the night before. That mania fuels an amphetamine like focus.

The upshot, if my hypothesis is true, is that waking up early would not produce similar gains if you did it everyday. Like the depressive who stays up all night, it may feel like you've discovered an intervention that will pay lasting gains. But if you were to actually make it part of your recurring lifestyle, the benefits would stop, and eventually the impact would work in reverse.

Along those lines that's probably why you naturally tend to stop conforming to that pattern after a few days. As acute sleep deprivation becomes chronic, you're most likely intuitively recognizing that the pattern has crossed over to the point of being counter-productive.

Lots of writers and software engineers note that their creative juices start flowing by evening extending late into the night - I think this phenomenon is closely related to the one described in the comment above.

Brian Timar:

sleep anecdote- In undergrad I had zero sleep before several major tests; also before quals in grad school. Basically wouldn't sleep before things I really considered important (this included morning meetings I didn't want to miss!). On such occasions I would feel:

- miserable, then
- absurd and in a good humor, weirdly elated, then
- Super Pumped™, and

really sharp when the test (or whatever) actually started.

Appendix: anecdotes about long-term sleep deprivation

I once tried to cheat sleep, and for a year I succeeded (strong peak-performance-sailing vibes):

In the summer of 2009, I was finishing the first—and toughest—year of my doctorate. ...

To keep up this crazy sleep schedule, I always needed a good reason to wake up the next morning after my 3.5-hour nighttime sleep. So before I went to bed, I reviewed the day gone past and planned what I would do the next day. I've carried on with this habit, and it serves me well even today.

But the Everyman schedule was reasonably flexible. Some days when I missed a nap, I simply slept a little more at night. There were also days when I couldn't manage a single nap, but it didn't seem to affect me very much the next day.

To the surprise of many, and even myself, I had managed to be on the polyphasic schedule for more than a year. But then came a conference where for a week I could not get a single nap. It was unsettling but I was sure I would be able to get back to sleeping polyphasic without too much trouble.

I was wrong. When I tried to get back into the schedule, I couldn't find the motivation to do it; I didn't have the same urgent goals that I had had a year ago. So I returned to sleeping like an average human.

James Gleck in *Chaos* on Mitch Feigenbaum:

In the spring of 1976 he entered a mode of existence more intense than any he had lived through. He would concentrate as if in a trance, programming furiously, scribbling with his pencil, programming again. He could not call C division for help, because that would mean signing off the computer to use the telephone, and reconnection was chancy. He could not stop for more than five minutes' thought, because the computer would automatically disconnect his line. Every so often the computer would go down anyway, leaving him shaking with adrenaline. He worked for two months without pause. His functional day was twenty-two hours. He would try to go to sleep in a kind of buzz, and awaken two hours later with his thoughts exactly where he had left them. His diet was strictly coffee. (Even when healthy and at peace, Feigenbaum subsisted exclusively on the reddest possible meat, coffee, and red wine. His friends speculated that he must be getting his vitamins from cigarettes.)

In the end, a doctor called it off. He prescribed a modest regimen of Valium and an enforced vacation. But by then Feigenbaum had created a universal theory.

[Ryan Kulp's experience with decreasing the amount of sleep by several hours:](#)

i began learning to code in 2015. since i was working full-time i needed to maximize after-hours to learn quickly. i experimented for 10 days straight... go to sleep at 4am, wake up at 8am for work. felt fine.

actually, the first 5-10 minutes of "getting up" after 3-4 hours of sleep sucks more than if i sleep ~8 hours. but after 15 mins of moving around, a shower, etc, i feel as if i slept 8 hours.

since then i've routinely slept 4-6 hours /day and definitely been more productive. i think if more people experimented for themselves and had the same "aha" moment i did (that you feel fine after the initial gut-wrenching "i slept too little" reaction), they'd get more done too.

This is a very good point that shows that: there's (1) how sleepy we feel when waking up and (2) how sleepy we feel during the day. (2) is probably more important but most people are focused on (1) and the implicit assumption is that poor (1) leads to (2) - which is unwarranted.

Appendix: how I wake up after 6 or less hours of sleep

[Nabeel Qureshi](#) writes:

you're combining two things here: (1) your brain is overpowered by the comfy soft temp-controlled bed (2) you're bored. they might both be right but i think you conflate them, and they're separate arguments. this is important bc i think the strongest counterargument to what you're saying is the classic experience of: you force yourself to wake up early (say 6), you have a project you're genuinely excited about (hence #2 is false), but when you sit down to work, you're tired and can't quite focus. in this scenario, i think your theory would say that i'm not really that excited about what i'm doing, because if i were (see video game argument) then i'd be awake. i'd disagree and say that the researcher should just go take a nap, and they'll probably be able to make more progress per hour than the extra hours they gain... trying to force yourself to do

something while underslept, subjectively, feels hellish. I'm sure you've had this experience - did you figure out a workaround?

It is completely true that if you are excited by a project but it's not super stimulating, it's still very easy to wake up after less than usual number of hours of sleep and feel sleepy and terrible. This is true for me as well. I found a solution to this: instead of heading straight to the computer, I first unload the clean plates from the dishwasher and load it with dirty plates. This activity is quite special in that it is:

1. Physical (includes lots of moving around physical objects to/from around the kitchen).
 - Why this matters: moving around wakes up the body much better than just sitting.
2. Mental (the objects are always in different places, the arrangement of them within the dishwasher is always somewhat different and you need to effectively solve a new spatial organization problem every day to load everything efficiently).
 - Why this matters: moving around in automatic pre-defined movements eventually results in the brain just performing these movements on autopilot without waking up.
3. Very moderate in effort (no lifting of heavy things, nothing that requires complex concentration).
 - why this matters: I and people I know tend to find intense physical activities right after waking up really unpleasant and somewhat nauseating.

In about 90% of the cases, 10 minutes later when I'm done with the dishwasher, I find that I'm fully awake and don't actually want to sleep anymore. In the remaining 10% of the cases, I stay awake and work until my wife wakes up and then go take a 20-minute nap under her watch (and take as many 20-minute naps as I need during the day, although I only end up taking a few naps a week and rarely more than one per day, unless I'm sick).

Appendix: Elon Musk on working 120 hours a week and sleep

[CNBC](#):

On Tesla's first-quarter earnings conference call in May, Musk referred to inquiries from Wall Street analysts as "boring, bonehead questions" and as "so dry. They're killing me." On the next earnings conference call in August, Musk said he was sorry for "being impolite" on the previous call.

"Obviously I think there's really no excuse for bad manners and I was violating my own rule in that regard. There are reasons for it, I got no sleep, 120 hour weeks, but nonetheless, there is still no excuse, so my apologies for not being polite on the prior call," Musk said.

Later in August, in conversation with the New York Times, Musk reported using prescription sleep medication Ambien to sleep.

"Yeah. It's not like for fun or something," Musk told Swisher Wednesday. "If you're super stressed, you can't go to sleep. You either have a choice of, like, okay, I'll have zero sleep and then my brain won't work tomorrow, or you're gonna take some kind of sleep medication to fall asleep."

Musk said he was working such insane hours to get Tesla through the ramp up in production for its Model 3 vehicle. "[A]s a startup, a car company, it is far more difficult to be successful than if you're an established, entrenched brand. It is absurd that Tesla is alive. Absurd! Absurd."

Appendix: Philipp Streicher on homeostasis, its relationship to mania/depression, and on other points I make

[Philipp \(@Cautes\):](#)

First, I wanted to share a way of thinking about some of your findings that builds on the idea of a homeostatic control system (brought to you from engineering via cybernetics). The classic example is a thermostat, which keeps temperature of a room close to a set point. Biology is quite a bit more messy than this, of course, but the body makes use of a plenty of feedback mechanisms to stay close to set points as well. You're right in pointing out that these set points don't need to be healthy though. For example, measured via EEG, PTSD patients have alpha power (which primarily modulates neural inhibition in frontal, parietal and occipital areas of the brain) set points far below that of healthy control groups. One way to deal with these suboptimal set points is to simply disrupt the system. Here's a model that makes this point nicely: imagine all possible brain state dynamics as a two-dimensional plane and place a ball on it which represents the current brain state space. As the ball moves, the brain dynamics change as well (in frequency, phase, amplitude - you name it). On the plane, you have basins that give stability to the brain state, and repellers in the form of hills, as well as random noise and outside interference which drives the ball into various directions. Sometimes the ball will get stuck in basins which are highly suboptimal, but they are deep enough that exploration of other set points is not possible. If the system is disrupted, the ball might get jolted out of its basin though, and be again able to fall into a more optimal position.

With that said, there's plenty of evidence that stability in itself (even within better basins) is suboptimal for perfect health, because contexts change. For example, people who are very physically healthy (athletes, for example), tend to have far greater variance in the time interval between individual heart beats (heart rate variability) than even the average person, and as the average person gets healthier, their heart rate variability increases as well. Basically, the body becomes more resilient by introducing a noise signal that produces chaotic fluctuations to homeostatic control mechanisms (controlled allostasis) and there are good reasons to think that this is true of psychological health as well.

Because of this, I think that you're right in suggesting that varying the amount of time you sleep is a good thing - especially if you're currently struggling with depression or mania. Not even necessarily because sleep per se is the culprit, but because it might dislodge a ball stuck in a suboptimal basin, so to speak. Depressed people tend to oversleep, people with mania tend to sleep too little, so steering in the opposite direction is only logical. For perfectly healthy people, sleep cycling is probably the best way to go - kind of a mirroring the logic of heart rate variability: introduce some noise to keep your body on your toes. It's just like fasting, working out, cold exposure, saunas, etc. - it's all about producing stressors on the body which stir up repair processes which keep you healthy (and biologically younger). I have done plenty of self-experiments with polyphasic 5-6 hour sleeping (similar to the approach studied by Stampi, who you mentioned), with no negative consequences. The main thing that makes it impractical is that intermittent napping is sometimes hard to combine with professional responsibilities and a social life.

As a side note, because you ask the question about why depressed people sleep longer, and people with mania sleep less, the answer to this is very likely highly multi-causal. With that said, I wanted to point out that depressed people generally exhibit excessive alpha activity in eyes-open waking states, which normally becomes more pronounced in people as they drift off to sleep (because of the neural inhibition function). We also have reason to believe that it mediates between BDNF and subclinical depressed mood, so

that's a link to something else you talk about in your article. As for mania, I haven't looked at this myself, but I remember hearing that it's almost a mirror image, with generally decreased synchronisation of slower oscillations and heightened faster rhythms, generally associated with greater arousal and wakefulness.

One last thing: as you point out, sleep is likely not required for memory retention. Any claim that sleep is about any specific cognitive function should be suspect on the principle that the phenomenon of sleep predates the development of organisms with brains - it can't have evolved specifically for something as high-level as memory retention. It's more likely about something more basic like general metabolic health.

Appendix: Jerome Siegel and Robert Vertes vs the sleep establishment

Time for the Sleep Community to Take a Critical Look at the Purported Role of Sleep in Memory Processing by Robert Vertes and Jerome Siegel (a reply to Walker claiming that the debate on memory processing in sleep is essentially settled):

The present 'debate' was sparked by an editorial by Robert Stickgold in SLEEP on an article in that issue by Schabus et al on paired associate learning and sleep spindles in humans

Regarding Stickgold's editorial, I was particularly troubled by his opening statement, as follows: "The study of sleep-dependent memory consolidation has moved beyond the question of whether it exists to questions of its extent and of the mechanisms supporting it". He then proceeded to cite evidence justifying this statement. Surprisingly, there was no mention of opposing views or a discussion of data inconsistent with the sleep-memory consolidation (S-MC) hypothesis. It seemed that the controversial nature of this issue should have at least been acknowledged, but apparently to do so would have undermined Stickgold's position that the door is closed on this debate and only the fine points need be resolved. ...

1. **By all accounts, sleep does not serve a role in declarative memory. As reviewed by Smith, with few exceptions, reports have shown that depriving subjects of REM sleep does not disrupt learning/memory, or exposure to intense learning situations does not produce subsequent increases in REM sleep.** Smith concluded: "REM sleep is not involved with consolidation of declarative material." The study by Schabus et al (see above) is another example that the learning of declarative material is unaffected by sleep. They reported that subjects showed no significant difference in the percentage of word-pairs correctly recalled before and after 8 hours of sleep. Or as Stickgold stated in his editorial [the editorial Vertes and Siegel are replying to], "Performance in the morning was essentially unchanged from the night before". It would seem important for Stickgold/Walker to acknowledge that the debate on sleep and memory has been reduced to a consideration of procedural memory - to the exclusion of declarative memory. If there are exceptions, they should note.
2. Several lines of evidence indicate that REM sleep is not involved in memory processing/consolidation - or at least not in humans. Perhaps the strongest argument for this is the demonstration that the marked suppression or elimination of REM sleep in individuals with brainstem lesions or on antidepressant drugs has no detrimental effect on cognition. A classic case is that of an Israeli man who at the age of 19 suffered damage to the brainstem from shrapnel from a gunshot wound, and when examined at the age of 33 he showed no REM sleep. The man, now 55, is a lawyer, a painter and interestingly the editor of a puzzle column for an Israeli magazine. Recently commenting on his 'famous' patient, Peretz Lavie stated that "he is probably the most normal person I know and one of the most successful

ones". There are several other well documented cases of individuals with greatly reduced or absent REM sleep that exhibit no apparent cognitive deficits. It would seem that these individuals would be a valuable resource for examining the role of sleep in memory. ...

In [Memory Consolidation Is Similar in Waking and Sleep](#) cited above, Siegel notes:

To critically evaluate this hypothesis [that sleep has a critical role in memory consolidation], we must take "interference" effects into account. **If you learn something before or after the experimenter induced learning that is being measured in the typical sleep-memory study, it degrades recall of the tested information.** For example if you tell a subject that the capital of Australia is Canberra and then allow the subject to have a normal night's sleep, there is a high probability that the subject will remember this upon awakening. If on the other hand you tell the subject that the capital of Australia is Canberra, the capital of Brazil is Brasilia, the capital of Canada is Ottawa, the capital of Iceland is Reykjavik, the capital of Libya is Tripoli, the capital of Pakistan is Islamabad, etc., it is much less likely the subject will remember the capital of Australia. The effect of proactive and retroactive interference is dependent on the temporal juxtaposition, complexity, and similarity of the encountered material to the associations being tested. Interference is a well-established concept in the learning literature [1-6]. **Under interference conditions, such as exist during sleep deprivation, subjects, by staying awake, necessarily interacting with the experimenter keeping them awake and experiencing the laboratory environment, will remember more than just the items that are presented. But they may be less able to recall the particular items the experimenter is measuring. This can lead to the mistaken conclusion that sleep is required for memory consolidation [7].**

[Fur Seals Suppress REM Sleep for Very Long Periods without Subsequent Rebound:](#)

Virtually all land mammals and birds have two sleep states: slow-wave sleep (SWS) and rapid eye movement (REM) sleep [1, 2]. After deprivation of REM sleep by repeated awakenings, mammals increase REM sleep time [3], supporting the idea that REM sleep is homeostatically regulated. ***Some evidence suggests that periods of REM sleep deprivation for a week or more cause physiological dysfunction and eventual death [4, 5]. However, separating the effects of REM sleep loss from the stress of repeated awakening is difficult [2, 6].** The northern fur seal (*Callorhinus ursinus*) is a semiaquatic mammal [7]. It can sleep on land and in seawater. The fur seal is unique in showing both the bilateral SWS seen in most mammals and the asymmetric sleep previously reported in cetaceans [8]. Here we show that **when the fur seal stays in seawater, where it spends most of its life [7], it goes without or greatly reduces REM sleep for days or weeks. After this nearly complete elimination of REM, it displays minimal or no REM rebound upon returning to baseline conditions.** Our data are consistent with the hypothesis that REM sleep may serve to reverse the reduced brain temperature and metabolism effects of bilateral nonREM sleep, a state that is greatly reduced when the fur seal is in the seawater, rather than REM sleep being directly homeostatically regulated. This can explain the absence of REM sleep in the dolphin and other cetaceans and its increasing proportion as the end of the sleep period approaches in humans and other mammals.

Appendix: more papers I found interesting

[Long-term moderate elevation of corticosterone facilitates avian food-caching behaviour and enhances spatial memory](#)

It is widely assumed that chronic stress and corresponding chronic elevations of glucocorticoid levels have deleterious effects on animals' brain functions such as

learning and memory. Some animals, however, appear to maintain moderately elevated levels of glucocorticoids over long periods of time under natural energetically demanding conditions, and it is not clear whether such chronic but moderate elevations may be adaptive. I implanted wild-caught food-caching mountain chickadees (*Poecile gambeli*), which rely at least in part on spatial memory to find their caches, with 90-day continuous time-release corticosterone pellets designed to approximately double the baseline corticosterone levels. Corticosterone-implanted birds cached and consumed significantly more food and showed more efficient cache recovery and superior spatial memory performance compared with placebo-implanted birds. Thus, contrary to prevailing assumptions, long-term moderate elevations of corticosterone appear to enhance spatial memory in food-caching mountain chickadees. These results suggest that moderate chronic elevation of corticosterone may serve as an adaptation to unpredictable environments by facilitating feeding and food-caching behaviour and by improving cache-retrieval efficiency in food-caching birds.

References

- Beersma DG, Van den Hoofdakker RH. Can non-REM sleep be depressogenic?. *Journal of affective disorders*. 1992 Feb 1;24(2):101-8.
- Bessone P, Rao G, Schilbach F, Schofield H, Toma M. The economic consequences of increasing sleep among the urban poor. *The Quarterly Journal of Economics*. 2021 Aug;136(3):1887-941.
- Consensus Conference Panel:, Watson, N.F., Badr, M.S., Belenky, G., Blwise, D.L., Buxton, O.M., Buysse, D., Dinges, D.F., Gangwisch, J., Grandner, M.A. and Kushida, C., 2015. Joint consensus statement of the American Academy of Sleep Medicine and Sleep Research Society on the recommended amount of sleep for a healthy adult: methodology and discussion. *Journal of Clinical Sleep Medicine*, 11(8), pp.931-952.
- Eckert A, Karen S, Beck J, Brand S, Hemmeter U, Hatzinger M, Holsboer-Trachsler E. The link between sleep, stress and BDNF. *European Psychiatry*. 2017 Apr;41(S1):S282-.
- Giese M, Unternährer E, Hüttig H, Beck J, Brand S, Calabrese P, Holsboer-Trachsler E, Eckert A. BDNF: an indicator of insomnia?. *Molecular psychiatry*. 2014 Feb;19(2):151-2.
- Goldschmied JR, Rao H, Dinges D, Goel N, Detre JA, Basner M, Sheline YI, Thase ME, Gehrman PR. 0886 Recovery Sleep Significantly Decreases BDNF In Major Depression Following Therapeutic Sleep Deprivation. *Sleep*. 2019 Apr;42(Supplement_1):A356-.
- Horne JA, Pettitt AN. High incentive effects on vigilance performance during 72 hours of total sleep deprivation. *Acta psychologica*. 1985 Feb 1;58(2):123-39.
- Kaiser J. More than half of high-impact cancer lab studies could not be replicated in controversial analysis. AAAS Articles DO Group. 2021;
- Kaplan RM, Irvin VL. Likelihood of null effects of large NHLBI clinical trials has increased over time. *PLoS one*. 2015 Aug 5;10(8):e0132382.
- Lyamin OI, Kosenko PO, Korneva SM, Vyssotski AL, Mukhametov LM, Siegel JM. Fur seals suppress REM sleep for very long periods without subsequent rebound. *Current Biology*. 2018 Jun 18;28(12):2000-5.
- Pravosudov VV. Long-term moderate elevation of corticosterone facilitates avian food-caching behaviour and enhances spatial memory. *Proceedings of the Royal Society of London. Series B: Biological Sciences*. 2003 Dec 22;270(1533):2599-604.

Turner EH, Matthews AM, Linardatos E, Tell RA, Rosenthal R. Selective publication of antidepressant trials and its influence on apparent efficacy. *New England Journal of Medicine*. 2008 Jan 17;358(3):252-60.

Rahmani M, Rahmani F, Rezaei N. The brain-derived neurotrophic factor: missing link between sleep deprivation, insomnia, and depression. *Neurochemical research*. 2020 Feb;45(2):221-31.

Riemann, D., König, A., Hohagen, F., Kiemen, A., Voderholzer, U., Backhaus, J., Bunz, J., Wesiack, B., Hermle, L. and Berger, M., 1999. How to preserve the antidepressive effect of sleep deprivation: A comparison of sleep phase advance and sleep phase delay. *European archives of psychiatry and clinical neuroscience*, 249(5), pp.231-237.

Seystahl K, Könnecke H, Sürütü O, Baumann CR, Poryazova R. Development of a short sleeper phenotype after third ventriculostomy in a patient with ependymal cysts. *Journal of Clinical Sleep Medicine*. 2014 Feb 15;10(2):211-3.

Shen X, Wu Y, Zhang D. Nighttime sleep duration, 24-hour sleep duration and risk of all-cause mortality among adults: a meta-analysis of prospective cohort studies. *Scientific Reports*. 2016 Feb 22;6:21480.

Shi G, Xing L, Wu D, Bhattacharyya BJ, Jones CR, McMahon T, Chong SC, Chen JA, Coppola G, Geschwind D, Krystal A. A rare mutation of β 1-adrenergic receptor affects sleep/wake behaviors. *Neuron*. 2019 Sep 25;103(6):1044-55.

Siegel JM. Memory Consolidation Is Similar in Waking and Sleep. *Current Sleep Medicine Reports*. 2021 Mar;7(1):15-8.

Sterr A, Kuhn M, Nissen C, Ettine D, Funk S, Feige B, Umarova R, Urbach H, Weiller C, Riemann D. Post-stroke insomnia in community-dwelling patients with chronic motor stroke: physiological evidence and implications for stroke care. *Scientific Reports*. 2018 May 30;8(1):8409.

Vertes RP, Siegel JM. Time for the sleep community to take a critical look at the purported role of sleep in memory processing. *Sleep*. 2005 Oct 1;28(10):1228-9.

Xie L, Kang H, Xu Q, Chen MJ, Liao Y, Thiagarajan M, O'Donnell J, Christensen DJ, Nicholson C, Iliff JJ, Takano T. Sleep drives metabolite clearance from the adult brain. *science*. 2013 Oct 18;342(6156):373-7.

Yetish G, Kaplan H, Gurven M, Wood B, Pontzer H, Manger PR, Wilson C, McGregor R, Siegel JM. Natural sleep and its seasonal variations in three pre-industrial societies. *Current Biology*. 2015 Nov 2;25(21):2862-8.

Youngstedt SD, Goff EE, Reynolds AM, Kripke DF, Irwin MR, Bootzin RR, Khan N, Jean-Louis G. Has adult sleep duration declined over the last 50+ years?. *Sleep medicine reviews*. 2016 Aug 1;28:69-85.

Butterfly Ideas

Or "How I got my hyperanalytical friends to chill out and vibe on ideas for 5 minutes before testing them to destruction"

Sometimes talking with my friends is like intellectual combat, which is great. I am glad I have such strong cognitive warriors on my side. But not all ideas are ready for intellectual combat. If I don't get my friend on board with this, some of them will crush an idea before it gets a chance to develop, which feels awful and can kill off promising avenues of investigation. It's like showing a beautiful, fragile butterfly to your friend to demonstrate the power of flight, only to have them grab it and crush it in their hands, then point to the mangled corpse as proof butterflies not only don't fly, but can't fly, look how busted their wings are.



You know who you are

When I'm stuck in a conversation like that, it has been really helpful to explicitly label things as butterfly ideas. This has two purposes. First, it's a shorthand for labeling what I want (nurturance and encouragement). Second, it explicitly labels the idea as *not ready for prime time* in ways that make it less threatening to my friends. They can support the exploration of my idea without worrying that support of exploration conveys agreement, or agreement conveys a commitment to act.

This is important because very few ideas start out ready for the rigors of combat. If they're not given a sheltered period, they will die before they become useful. This cuts us off from a lot of goodness in the world. Examples:

- A start-up I used to work for had a keyword that meant “I have a vague worried feeling I want to discuss without justifying”. This let people bring up concerns before they had an ironclad case for them and made statements that could otherwise have felt like intense criticism feel more like information sharing (they’re not asserting this will definitely fail, they’re asserting they have a feeling that might lead to some questions). This in turn meant that problems got brought up and addressed earlier, including problems in the classes “this is definitely gonna fail and we need to make major changes” and “this excellent idea but Bob is missing the information that would help him understand why”.
 - This keyword was “FUD (fear, uncertainty, doubt)”. It is used in exactly the opposite way in cryptocurrency circles, where it means “you are trying to increase our anxiety with unfounded concerns, and that’s bad”. Words are tricky.
- [Power Buys You Distance From The Crime](#) started out as a much less defensible seed of an idea with a much worse explanation. I know that had I talked about it in public it would have caused a bunch of unproductive yelling that made it harder to think because I did and it did (but later, when it was ready, intellectual combat with John Wentworth improved the idea further).
- The entire genre of “Here’s a cool new emotional tool I’m exploring”
- The entire genre of “I’m having a feeling about a thing and I don’t know why yet”

I've been on the butterfly crushing end of this myself- I'm thinking of a particular case last year where my friend brought up an idea that, if true, would require costly action on my part. I started arguing with the idea, they snapped at me to stop ruining their dreams. I chilled out, we had a long discussion about their goals, how they interpreted some evidence, and why they thought a particular action might further said goals, etc.

A week later all of my objections to the specific idea were substantiated and we agreed not to do the thing- but thanks to the conversation we had in the meantime, I have a better understanding of them and what kinds of things would be appealing to them in the future. That was really valuable to me and I wouldn't have learned all that if I'd crushed the butterfly in the beginning.

Notably, checking out that idea was fairly expensive, and only worth it because this was an extremely close friend (which both made the knowledge of them more valuable, and increased the payoff to helping them if they'd been right). If they had been any less close, I would have said “good luck with that” and gone about my day, and that would have been a perfectly virtuous reaction.

I almost never discuss butterfly ideas on the public internet, or even 1:many channels. Even when people don't actively antagonize them, the environment of Facebook or even large group chats means that people often read with half their brain and respond to a simplified version of what I said. For a class of ideas that live and die by context and nuance and pre-verbal intuitions, this is crushing. So what I write in public ends up being on the very defensible end of the things I think. This is a little bit of a shame, because the returns to finding new friends to study your particular butterflies with is so high, but ce la vie.

This can play out a few ways in practice. Sometimes someone will say “this is a butterfly idea” before they start talking. Sometimes when someone is being inappropriately aggressive towards an idea the other person will snap “will you please stop crushing my butterflies!” and the other will get it. Sometimes someone will overstep, read the other's facial expression, and say “oh, that was a butterfly, wasn't it?”. All of these are marked improvements over what came before, and have led to more productive discussions with less emotional pain on both sides.

On Bounded Distrust

Response To (Scott Alexander): [Bounded Distrust](#)

Would that it were that simple.

There is a true and important core idea at the center of Bounded Distrust.

You can (and if you are wise, you often do) have an individual, institution or other information source that you *absolutely do not trust* to reliably tell the truth or follow its own explicit rules. Yet by knowing the implicit rules, and knowing the incentives in place and what the consequences would be if the source was lying to various extents, you can still extract much useful information.

Knowing what information you can and can't extract, and what claims you can trust from what sources in what contexts to what extent, is a vital life skill.

It is also a difficult and often an [anti-inductive](#) skill. Where there is trust, there is the temptation to abuse that trust. Each person has a unique set of personal experiences in the world and samples a different set of information from various sources, which then evolves one's physical world models and one's estimates of trustworthiness in path dependent ways. Making real efforts from your unique epistemic perspective, will result in a unique set of heuristics for who you can and cannot trust.

Perspectives can become less unique when people decide to merge perspectives, either because they trust each other and can trade information on that basis, or because people are conforming and/or responding to social pressure. In extreme cases large groups adopt an authority's stated heuristics wholesale, which that authority may or may not also share.

Scott's model and my own have much in common here, but also clearly have strong disagreements on how to decide what can and cannot be trusted. A lot of what my weekly Covid posts are about is figuring out how much trust we can place where, and how to react when trust has been lost.

This all seems worth exploring more explicitly than usual.

I'll start with the parts of the model where we agree, in list form.

1. None of our institutions can be trusted to always tell the truth.
2. However, there are still rules and associated lines of behavior.
3. Different rules have different costs associated with breaking them.
4. These costs vary depending on the details, and who is breaking what rule.
5. This cost can, in some cases, be so high as to be existential.
6. In some situations, this cost is low enough that statements cannot be trusted.
7. In some situations, this cost is high enough that statements can be trusted.
8. Often there will be an implicit conspiracy to suppress true information and beliefs, but the participants will avoid claiming the information is false.
9. Often there will be an implicit conspiracy to spread false information and beliefs, but the participants will avoid explicitly claiming false information.
10. This conspicuous lack of direct statements is often very strong evidence.
11. [The use of 'no evidence' and its synonyms is also strong evidence.](#)

12. There will sometimes be ‘bounded lying’ where the situation is painted as different than it is but only by a predictable fixed amount. If you know the rules, you can use this to approximate the true situation.

The difference is that Scott seems to think that the government, media and other authority figures continue mostly to play by a version of these rules that I believe they *mostly used to* follow. He doesn’t draw any parallels to the past, but his version of bounded distrust reads like something one might plausibly believe in 2015, and which I believe was largely the case in 1995. I am confused about how old the old rules are, and which ones would still have held mostly true in (for example) 1895 or in Ancient Rome.

Whereas in 2022, after everything that has happened with the pandemic and also otherwise, I strongly believe that the trust and epistemic commons that existed previously have been burned down. The price of breaking the old rules is lower, but it is more than that. The price of being viewed as actually following the old rules is higher than the cost of not following them, in addition to the local benefits of breaking the old rules. Thus the old rules mostly are not followed.

The new rules are different. They still bear some similarities to the old rules. One of the new rules is to pretend (to pretend?) to be following the old rules, which helps. The new rules are much less about tracking physical truth and much more about tracking narrative truth.

It seems useful to go through Scott’s examples and some obvious variants of them as intuition pumps, but first seems worth introducing the concept of the One Time, and also being clear about what this post *doesn’t* discuss due to time and length constraints, since that stuff is very important.

Bounded Discussion (Or: Why Can Wait)

This is the long version of this post due to lack of time to write a shorter one. I do hope at some point to write a shorter version.

Thus, this post is already extremely long, so it doesn’t have additional space to get much into my model of *why* much of this is the case.

Here are *some* of the things I am *conspicuously* excluding due to length and time.

I’m excluding all [discussion of simulacra levels](#), and [all discussion of moral mazes](#) or even [motive ambiguity](#), or the dynamics of implicit conspiracies, despite them being important to the underlying dynamics. I’m excluding all the reasons why there is pressure to *break the rules* as much as possible and be seen to be doing so, and why this pressure is currently historically high and increasing over time along with pressures to visibly reverse all principles, values and morality.

Thus I’m not ‘putting it all together’ in an important sense. Not yet.

I’m excluding all discussion of what the Narrative actually is, how it gets constructed and decided upon, what causes it to update and to what extent it responds to changes in the physical world.

I’m excluding all discussion of why there exists an Incorrect Anti-Narrative Contrarian Cluster (ICC), or a Correct Contrarian Cluster (CCC), or how their dynamics work in

terms of what gets included or excluded, or what pushes people towards or away from them.

I'm excluding most of the necessary discussion of how one evaluates a particular source to decide how bounded one's distrust in that particular source should be and in which particular ways, and how I identify sources I can mostly or entirely trust or ones which are less trustworthy than their basic identify would suggest.

I'm excluding discussion about how to *elicit* truth from by-default untrustworthy sources, if given the opportunity to interact with them, which is often possible to varying degrees.

I'm excluding a bunch of synthesis that requires more careful simmering and making into its own shorter post.

I'm excluding a bunch of other things too. This is quite the rabbit hole to go down.

Now, on to the One Time.

One Time

There's a lot of great tricks that only work once. They work because it's a surprise, or because you're spending a unique resource that can't be replenished.

The name here comes from poker, where players jokingly refer to their 'One Time' to get lucky and hit their miracle card. Which is a way of saying, *this is the one that counts*.

That's the idea of the One Time. This is the high leverage moment. It will give away your secret strategy, show your opponent their weakness, blow your credibility, use all your money, get you fired, wreck the house or cash in that big favor. The rules will adjust and it won't work again.

Or alternatively, *if it succeeds* you get your One Time back, but damn it This Had Better Work or there will be hell to pay. [You come at the king, you best not miss.](#)

Maybe that's acceptable. Worth It. Then do what you have to do.

That's one way of thinking about the price one has to pay for breaking some of these rules. Would this be slight annoyance? Or would you be cashing in your One Time?

Shooting at Yankee Stadium

Scott frames this as you being a liberal and thus lacking trust in Fox as a source, but it's important to note that *this does not matter*. Either Fox News is trustworthy in a given situation, or it is not. MSNBC is also either trustworthy in a given situation to a given degree, or it is not. Your views on social and economic policies, and which party you want in power to what degree, should not matter. The exception is if the *reason* you are on one side or the other is that you believe one side's sources are more honest, but if that's true then you're a liberal because you don't trust Fox News, rather than not trusting Fox News because you're a liberal.

Anyway, this is his first example:

One day you're at the airport, waiting for a plane, ambiently watching the TV at the gate. It's FOX News, and they're saying that a mass shooter just shot twenty people in Yankee Stadium. There's live footage from the stadium with lots of people running and screaming.

Do you believe this?

Yes, of course I believe it. In fact it's rather overdetermined. Why?

1. Clear physical fact claims, specific details.
2. If false, will be known to be false quickly and clearly.
3. If caught getting this wrong, price would be high relative to stakes.
4. Admission against interest (could also say poor product-market fit).
5. Live footage from the stadium.
6. They are not in the habit of this type of lie.

The *combination* of these factors is very strong, and in the absence of counterevidence I would treat this as true with probability of essentially $(1 - \epsilon)$.

I agree with Scott that deep fakes of live events are beyond the reasonable capabilities of Fox News or any other similar organization at this time. And also that even if they could do them, the price of getting caught doing so would be very high, even higher than the already high price of being seen getting this wrong. So the live footage alone makes me believe whatever I see on the live footage.

I would only doubt that if the stakes involved were somehow high enough that Fox News would plausibly be cashing in their One Time (e.g. if #3 was false, because the value at stake rivaled the potential price).

Note of course that the live footage doesn't always mean what it looks like it means. It can and will be framed and edited to make it look like they want it to look, and anyone interviewed might be part of the production. It doesn't automatically imply a mass shooting. But you can trust the literal evidence of your senses.

If it was Yankee Stadium *without* live footage, that lack of footage would be highly suspicious, because there should be cameras everywhere and Fox should be able to get access. I'd wonder what was up. But let's say we move this to a place without such cameras, so it's not suspicious, or otherwise we don't have footage that actually proves that the shootings happened (and for whatever reason it's not suspicious that we lack this). Are we still good?

Yeah, we're still good. It's still reporting physical facts with specific details, in a way that if anything goes directly *against* Fox's vested interests. There's no reason to lie.

What if in addition to removing the live footage, it was MSNBC or CNN instead, so there was a clear reason to claim there was a mass shooting but the situation is otherwise unchanged?

Now I notice that this is a sufficiently *combination* of missing factors that I'm moving from numbers like $(p = 1 - \epsilon)$ to something more like $p \sim 0.95$. They *could* make a mistake here, they have *reason* to make a mistake here, they're in the *habit* of calling things mass shootings whenever possible. The *price* for getting this wrong isn't zero, but the mainstream media is good at memory holing its 'mistakes' of this type and isn't trying to be super reliable anymore.

They are *in the habit* of this kind of lie, of finding ways to claim there are lots of mass shootings all the time, and characterizing everything they can as a mass shooting, so #6 also does not apply, although there would still be *something* that their source was claiming had happened – they wouldn't as of yet be willing to use this label if none of their sources were saying bullets were involved or that anyone had come to harm.

It's *probably* still a mass shooting, but if my life depends on that being true, I'm going to double check.

Scott's next hypothetical:

Fox is saying that police have apprehended a suspect, a Saudi immigrant named Abdullah Abdul. They show footage from a press conference where the police are talking about this. Do you believe them?

Once again, yes, of course. This is no longer an admission against interest, but I notice this is an *actual* red line that won't be crossed. The police either apprehended a suspect named Abdullah Abdul from Saudi Arabia or they didn't, this can be easily verified, and there *will* be a price very much not worth paying if this is claimed in error. There is a strong habit of not engaging in false statements of this type.

If this was more speculative, they would use particular weasel words like 'believed to (be/have)' at which point all bets aren't quite off but the evidence is not very strong. If the weasel words aren't there, there's a reason.

However, I don't agree with this, and even more don't agree with the sign-reversed version of it (e.g. flop MSNBC for FOX and reverse all the facts/motivations accordingly):

It doesn't matter at all that FOX is biased. You could argue that "FOX wants to fan fear of Islamic terrorism, so it's in their self-interest to make up cases of Islamic terrorism that don't exist". Or "FOX is against gun control, so if it was a white gun owner who did this shooting they would want to change the identity so it sounded like a Saudi terrorist". But those sound like crazy conspiracy theories. Even FOX's worst enemies don't accuse them of doing things like this.

This very much *does not* sound like a crazy conspiracy theory. It is not crazy. Also it would not be a conspiracy. It would be some people making some stuff up, only in locally noticeably more brazen ways than those previously observed, and which we thus think is *unlikely*. But if someone came into Scott's office and said 'I think FOX's story today about that Saudi terrorist is importantly false' then it would be a mistake to suggest therefore putting this person on medication or asking them to go to therapy.

Of course it matters that FOX is biased and would very much like to make up a case of Islamic terrorism. FOX makes up cases of Islamic terrorism, the same way that MSNBC shoves them under a rug. And my lord, FOX would *totally love* to change the identity so it sounded like a Saudi terrorist. Of course they would. And MSNBC would love to make it sound like it was a white gun owner.

Before the identity is known, MSNBC and friends will run stories that assume *of course* it is a white gun owner, while FOX and friends will run stories that assume *of course* it is an Islamic terrorist. And they will hold onto those assumptions until the *last possible moment* when it would be too embarrassing not to fold, in the hopes of leaving the right impression (for their purposes) with as many people as possible, and to signal

their loyalty to their narrative model of the world. And then they will insist they didn't say the things they previously said, and they will definitely insist they certainly haven't repeated this pattern dozens of times.

The question is, does adding the detail of the police identifying the suspect sufficiently over-the-line that this is *insufficient* to make the actions in question plausible? With these details, my answer is yes in the central sense of there being an apprehended suspect named Abdullah Abdul from Saudi Arabia.

Whereas on MSNBC, they're probably whistling and *pretending not to notice* this person's name and origin because they're suddenly not important, and having experts on saying things like 'we have no idea what caused this incident, who can know, but we do know that there are so many more shootings here than any other country.'

Now flip it again, and suppose the suspect was a white gun owner. Fox will keep talking about the threat of Islamic terrorism and *pretend not to notice* the person was white, and probably expound upon the various FOX-friendly potential motivations and histories that *could* be involved long after they're no longer remotely plausible.

Now imagine the person in question was *both*, and was a white person who happened to be born in Saudi Arabia, and whose name (whether or not it was given at birth) was Abdullah Abdul, and watch two *completely disjoint sets of facts* get mentioned.

But, you say. But! They still wouldn't outright say the fully false things here. There are rules, you say. Everyone involved is distorting everything in sight but there's still this big signpost where they say 'the police have apprehended a suspect named X with characteristics Y' and you know X is the suspect's name, and you *probably* know they have characteristics Y depending on how slippery that could potentially be made.

And yes, you're probably right. Last time I checked, they *do* have a red line there. But there's a bunch of red lines I thought they had (and that I think previously they did have) that they've crossed lately, so how confident can we be?

Scott says this:

And there are other lines you don't cross, or else you'll be the center of a giant scandal and maybe get shut down. I don't want to claim those lines are objectively reasonable. But we all know where they are. And so we all trust a report on FOX about a mass shooting, even if we hate FOX in general.

Scott links to Everybody Knows to indicate this is a 'the savvy know this and then treat it like everyone knows.' But the savvy are necessarily a subset, and not all that large a subset at that. Not only does everyone very much not know this, I don't even know this.

I have a *general sense* of where those lines seem to be, but they seem to be different than where the lines were five years ago, which in turn is different from thirty years ago. I am not confident I have them located correctly. I am *damn sure* that very far from everybody knows even that much with any confidence, and that those who think they are damn sure often strongly disagree with each other.

I *don't* expect this example to rise to anything like the level where FOX might get shut down and I'd expect it to be forgotten about within a few weeks except maybe for the occasional 'remember when FOX did X' on Twitter. They'll claim they made a mistake and got it wrong and who are you to say different and why should we believe your

biased opinion? That seems so much more likely to me than that this suddenly becomes a huge deal.

The reason I still believe FOX (or MSNBC in reverse) in this spot is because it's still not something they're in the habit of doing, and it's still a dumb move strategically to choose this spot to move expectations in this way, in ways they can understand intuitively, and mostly that it *feels* like something that *will feel to them* like something they shouldn't do. It doesn't pattern match well enough to the places where outright lies have already happened recently. Right now. For now.

Yet, for all our explicit disagreements, I expect Scott *in practice* to be using almost the same heuristics I am using here if such events were to happen, with the difference being that I think Scott should be adjusting more for recent declines in deserved trust, and him likely thinking I'm adjusting too far.

Lincoln and Marx

I'm going to first deal with the Lincoln and Marx example, then with the 2020 election after, although Scott switches back and forth between them.

[Here's a Washington Post article](#) saying that Abraham Lincoln was friends with Karl Marx and admired his socialist theories. It suggests that because of this, modern attacks on socialism are un-American.

[Here is a counterargument](#) that there's no evidence Abraham Lincoln had the slightest idea who Karl Marx was.

I find the counterargument much more convincing. Sometimes both the argument and counterargument describe the same event, but the counterargument gives more context in a way that makes the original argument seem calculated to mislead. I challenge you to read both pieces without thinking the same.

A conservative might end up in the same position vis-a-vis the *Washington Post* as our hypothetical liberal and FOX News. They know it's a biased source that often lies to them, but how often?

So both sides are often lying, but with some conditions under which a given statement can still be trusted. The question is what conditions still qualify.

So before looking at the counterargument, we can start with the easy observation that the headline is *definitely* at least a claim without evidence, which I would consider in context to be lying. Scott excuses this by saying that headline writers are distinct from article writers, and make stuff up, and everybody knows this and it's fine. Anything in a headline that isn't a tangible specific fact is complete rubbish.

The body of the article is a real piece of work. I didn't need to see the counterargument to know it stinks, only to know exactly *how much* it stinks. It is doing the association dance, the same one used when someone needs to be canceled. Other than being about someone long dead, and that the author thinks socialism is good actually, this seems a lot like [what The New York Times did to Scott Alexander](#), drawing the desired associations and implications by any means technically available, and because there was nothing there, being made of remarkably weak sauce.

Here Lincoln is ‘surrounded by’ a certain kind of person, and someone is that kind of person if they ‘made arguments’ that are of the type that a person of that point of view would make. I totally noticed that the argument that Lincoln was reading Marx was that he was a columnist in a newspaper Lincoln was reading, which is like saying I was as a child a reader of William Safire because I read the New York Times. The ‘exchanged letters’ thing where Lincoln wrote back a form letter I can’t say for sure I would have picked up on on my own, but I like to hope so. The clues are all there.

That’s the thing. The clues are all there. This is *transparent obvious bullshit*.

It’s still easy to not spot the transparent obvious bullshit. When one is reading casually or quickly, it’s a lot easier to do a non-literal reading that will effectively lie to you, than the literal reading that won’t. Not picking up on (or noticing in a conscious and explicit way that lets you reject them) the intended insinuations requires effort. And despite the overall vibe of the post being transparent enough to me that it would trigger a ‘only a literal reading of this will be anything but bullshit,’ it was less transparent to others – Scott said in a comment to a draft of this post that he’s not confident he would have sufficiently noticed if he’d seen only the original but not the rebuttal.

The *direct quotes* of Lincoln here are interesting. They do have quite the echo to things Marx said. And to things many others of that era said who had very different beliefs. They also make perfect sense if you interpret them as ‘you should want the slaves to be freed,’ which is the obvious presumed context when I read them, and which was then confirmed by the context later provided by the counterargument. Which also seems to include such lines as:

“Capital,” Lincoln explained, “has its rights, which are as worthy of protection as any other rights.”

They also are missing the thing that makes a socialist a socialist, which is to declare that we should find the people with the stuff, point guns at them, and take their stuff. It doesn’t even quote him saying this *about slaves*, and he’s the one who freed the slaves, so it seems like a strange omission. In this type of agenda-pushing, one can safely assume that if there was better material available it would have been used.

The counterargument misunderstands what is going on here.

Brockell badly misreads her sources and reaches faulty conclusions about the relationship between the two historical contemporaries. Contrary to her assertion, there is no evidence that Lincoln ever read or absorbed Marx’s economic theories. In fact, it’s unlikely that Lincoln even knew who Karl Marx was, as distinct from the thousands of well-wishers who sent him congratulatory notes after his reelection.

There’s the fact that *technically* no one said Lincoln read Marx’s economic theories but that’s not the point here. Brockell did not misread anything. Brockell looked for words that could be written to give an impression Brockell wished to convey while not crossing the red line of saying definitively false things of the wrong type, and Brockell found the best such words that could be found. There are no ‘faulty conclusions’ here, there are only implausible insinuations.

Anyway, yes, the rebuttal is deeply convincing, and the fact that the original made it into the Washington Post *should be* deeply embarrassing. Yet it was not. Scott notes that it was not, that everyone forgot about it. Scott seemingly thinks *not only* that the Washington Post will pay zero price for doing this, but that *this was entirely*

predictable. As a ‘human interest’ story, in his model, no one is checking for such obvious hackery or caring about it, it’s par for the course, you should expect to see ‘we can’t know for sure that wet ground causes rain, but we do know that there’s a strong correlation, and where wet ground you can usually look up and see the rain coming down’ and who cares, it’s not like it matters whether the rain caused the wet ground or the other way around, that’s a human interest story.

There’s also the question of *whether this story is lying or not*. Scott seems to be trying to have it both ways.

First off, there’s the common sense attitude that the Marx/Lincoln article is *of course* lying. But the claim is then that this is because the questions in it are not to be taken seriously, and trust only matters when questions are sufficiently serious.

Then there’s the thing where the article didn’t *technically* lie aside from the headline. Which is true.

It’s hard for a naïve person to read the article without falsely concluding that Marx and Lincoln were friends. But the article *does* mostly stick to statements which are literally true.

I don’t think it’s mostly? I think the statements are each *literally* true. It’s more like it’s full of insinuation and non-sequiturs. This paragraph, for example, is *all completely true* aside from the questionable ‘was surrounded by socialists’ but also is also *completely obvious nonsense*. It gives the impression that conclusions should be drawn without actually justifying those conclusions at all, which is classic.

President Trump has added a new arrow in his quiver of attacks as of late, charging that a vote for “[any Democrat](#)” in the next election “is a vote for the rise of radical socialism” and that Rep. Alexandria Ocasio-Cortez (D-N.Y.) and other congresswomen of color are “[a bunch of communists](#).” Yet the first Republican president, for whom Trump has expressed admiration, was surrounded by socialists and looked to them for counsel.

What are the potential outright falsehoods?

There’s that line about ‘surrounded by socialists’ above. The only evidence given is that there were a few people around Lincoln who expressed some socialist ideas, and who encouraged him to free the slaves. That doesn’t seem like it clears the bar on either ‘socialist’ or ‘surrounded.’ There are two socialists referenced, one of whom ran a Republican newspaper, supported him, and then investigated generals on his behalf, none of which has much to do with socialism. It’s no surprise that Lincoln ‘eagerly awaited’ dispatches about his generals, since his generals were one of his biggest issues. The other also ran a newspaper. It’s almost as if someone who wanted to run for office decided to become friends with the people who had access to printing presses. Smart guy, that Lincoln.

And there’s a bunch of statements like this. They seem more right than wrong, but not *quite wrong enough* to be lies.

There’s this:

If you think that sounds like something Karl Marx would write, well, that might be because Lincoln was regularly reading Karl Marx.

This is *highly misleading* in the sense that ‘regularly reading Karl Marx’ refers to his Crimea War dispatches in a newspaper, which he in turn *might or might not* have been doing, but technically that still counts. The question is whether the *logical implication* here counts as lying, since if you know the details it’s obvious that this could not have been why Lincoln wrote what he wrote.

Scott claims ‘the Marx article got minimal scrutiny’ but it manages to *very carefully* follow the correct exact pattern, and predictably got a bunch of scrutiny afterwards. I don’t buy it.

So my conclusion is that the article is intentionally misleading, a piece of propaganda designed to be obviously bullshitting in order to push a political agenda and make it clear you are willing to engage in obvious bullshit to support a political agenda.

But it’s *bullshit*, and isn’t *lying*, except for the headline. It follows The Rules, the Newspaperman’s Code that says that you can’t print known-to-be-technically-false things.

Human Interests

That leads to me getting confused by this.

Finally, the Marx thing was intended as a cutesy human interest story (albeit one with an obvious political motive) and [everybody knows](#) cutesy human interest stories are always false.

It could be reasonably said that [everybody knows](#) cutesy human interest stories are warped narratives at best and often centrally false, designed to give the desired impression and support the desired narrative. The post about rescuing that cat stuck in a tree is *either* going to talk about the *dark underbelly of shady cat rescuers* or else it’s going to be a heartwarming story about how a cute child got their kitty back. What it isn’t going to be is fair and balanced.

You can call this a ‘cutesy human interest story’ if you come from a background where being socialist is obviously great, but even then I don’t buy it because the purpose of this is to be used as ammunition in within-ingroup arguments to try and show one’s adherence to party lines. It’s not to try and convince any outgroup members because, as Dan Quayle famously put it and is quoted later in Scott’s post, no one was fooled.

Such people gave The Washington Post clicks, as did Scott here. Author showed their loyalties and ability to produce viral content of similar nature. Missions accomplished.

But the question I have is: **What makes the rules observed here different from the rules elsewhere?**

My answer to that is **nothing**. The rules are the same.

This is *exactly* the level of misleading one should expect, *at a minimum*, on a ‘how and in which way do Very Serious People want me to be worried this week about Covid-19.’ Or on a post about how an election (was / was not) stolen. This is *exactly* the level of misleading I expect *any time there is a narrative and an interest in pushing that narrative*.

In fact, I'd call this an excellent example of where the line used to be. The line used to be exactly here. You could do this. You couldn't do more.

The difference is that people are increasingly doing somewhat more than this. That's why we had to go through the steps earlier with the hypothetical shootings at Yankee Stadium. If 2012-media from any side tells me there's a mass shooting at Yankee Stadium, I believe them, full stop, we don't need the other supports. That's specific enough. Today, it's *not* enough, and we need to stop and think about secondary features.

It is often said that if you read an article in a newspaper about the field you know best it will make statements that are about as accurate as 'wet ground causes rain,' and you should then consider that maybe this isn't unique to the field you know best. That certainly matches my experience, and that's when there *isn't* an obvious narrative agenda involved. When there is, it's a lot worse.

Scott's attempt to draw the distinction that expert historians *specifically into Marx and Lincoln* are not known to be saying nice things about this article feels like ad hoc special pleading, a kind of motte/bailey on what contextually counts as an expert. It also isn't relevant, because 'praise' is not vouching even for its not-outright-lying status let alone its not-lying-by-implication status. Under the model, 'praise' is unprincipled, cannot be falsified, and thus doesn't imply what Scott is suggesting it does, and mostly is only evidence of what is in the Narrative.

Scott notices that he never expected any of this to check out under scrutiny, because stories like this are never true, and certainly there were overdetermined contextual clues to allow that sort of conclusion even before the takedown. With the takedown, it's trivial.

The 2020 Election

A conservative might end up in the same position vis-à-vis the *Washington Post* as our hypothetical liberal and FOX News. They know it's a biased source that often lies to them, but how often?

[Here's a Washington Post article](#) saying that the 2020 election wasn't rigged, and Joe Biden's victory wasn't fraudulent. In order to avoid becoming a conspiracy theorist, the conservative would have to go through the same set of inferences as the FOX-watching liberal above: this is a terrible news source that often lies to me, but it would be surprising for it to lie *in this particular case in this particular way*.

I think smart conservatives can do that in much the same way smart liberals can conclude the FOX story was real. The exact argument would be something like: the Marx article got minimal scrutiny. A few smart people who looked at it noticed it was fake, three or four people wrote small editorials saying so, and then nobody cared. The 2020 election got massive scrutiny from every major institution.

To be safe, I'll reiterate up front that **I am very confident the 2020 election was not rigged**. But I didn't get that confidence because liberal media sources told me everything was fine, I got it because I have a detailed model of the world where there's lots of strong evidence pointing in that direction. That and the stakes involved are why I broke my usual no-unnecessary-politics rules [in the post after the election was clearly decided](#) to be very explicit that Biden had won the election - it was a form

of cashing in one's One Time in a high-leverage moment, bending one's rules and paying the price.

As I write this, I haven't yet looked at the WaPo article so I can first notice my expectations. My expectation is that the WaPo article will have a strong and obvious agenda, and that it will be entirely unconvincing to anyone who hadn't already reached the conclusion that the 2020 election wasn't rigged, and will primarily be aimed at giving people a reference with which to feel smug about the stupid people who were 'fooled by the Big Lie' and think the 2020 election was rigged.

Notice that Scott's argument rests here on the *difference* between the election article and the Marx article. The Marx article *should not be believed*. But I notice that *I expect both articles to be following the same standards of evidence and honesty*. Whoops.

Enough preliminaries. Time to click and see what happens.

We can start with the headline. As we've established, *the headline is always bullshit*.

Guess what? There (still) wasn't any significant fraud in the 2020 presidential election.

So that's a *really strange* turn of phrase, isn't it? That still?

I mean, what would it mean for there to have not to *have been* fraud in the 2020 presidential election *at some point in the past*, looking back on the election, but for that to *have changed*, and there now to *have been* fraud that previously had not been?

Either there was fraud or there wasn't fraud. There's no way for that answer to change retroactively, unless the fraud took place in the interim, which isn't anyone's claim. So the mentality and model behind this headline is saying that *whether there was fraud* is somehow an *importantly different* claim than *whether or not someone did a fraudulent thing at the time*.

Instead, it's about *what the current narrative is*. The current narrative is that there wasn't fraud. The past narrative is that there wasn't fraud. Thus, there (still) wasn't any fraud, because 'there was fraud' means 'the narrative contains there being fraud.'

One can make claims about what is or is not in the narrative, under this lexicon, but there isn't an obvious combination of words that says whether or not someone did a physical act, only whether or not someone is generally said to have committed that act.

In other words, under this system, if I ask 'was there significant fraud in the 1960 presidential election?' I am asking whether the *narrative says* there was such fraud. And therefore the answer could be 'no' one day and 'yes' the next and then go back to 'no' based on what those who control the narrative prefer.

More charitably, one could interpret this as 'there is still no evidence for' (which is always false, there's never no evidence of anything) or 'there is still overwhelming evidence against' (which is both stronger and has the benefit of being true) and having been cut down because headlines have limited space, and that I'm reading too much into this.

I don't think so. The headline could have read "Evidence Still Overwhelmingly Says No Significant Fraud in 2020 Election" and been shorter. This was a choice. I think this headline is smug and has the asshole nature and makes it clear that this is how words are supposed to work and that none of this is an accident.

Let us begin.

It's been more than a year since the 2020 presidential election ended according to the calendar, though, according to the guy who clearly and unquestionably lost that election, Donald Trump, things are still up in the air. For 400 days, Trump has been promising sweeping evidence of rampant voter fraud in that election. It's eternally just around the corner, a week away. Two. It's his white whale and his Godot. It's never secured; it never arrives.

Yeah, that's all going to get past a fact checker as defensible things to say, but: Who is the intended audience here? Who is this trying to inform? If you are reading this while previously believing the election was stolen, do you keep reading? Given the headline, what's the chance you even got that far?

The entire article is written like this, with the *baseline assumption* that Trump is lying in bad faith and that claims of fraud are illegitimate.

Let's push through the fact that the whole thing has the asshole nature and has no interest in providing anything but smugness while pretending to be #Analysis, and look at what the *actual claims* are that one might 'have to be a conspiracy theorist' not to believe, since that's the core question, except man they make it hard.

Yet there he was, offering the same excuse once again when asked by the Associated Press. The occasion was [AP's exhaustive assessment](#) of the 2020 election in which they uncovered fewer than 500 questionable ballots. Questionable! Not demonstrably fraudulent, but questionable. But **Trump, never bound to reality**, waved it away.

If you're going to put lines like 'Trump, never bound to reality' into your statement, it's *really hard* to complain that people on the other side aren't viewing you as a credible source. You're spouting obvious nonsense for the sole purpose of delivering snappy one-liners and then wondering why those who think the target of that putdown should be in the White House aren't updating on your factual claims.

I mean, they end on this:

On the other hand, we have a guy who was documented as having said false things [tens of thousands of times](#) while serving as president continuing to insist that proof of wide-scale fraud is just around the corner.

But if you asked me to find *tens of thousands* of times the Washington Post has said that which was not, via a similar standard, do you think it would be *hard*?

We're agreed that they lie all the time. And Scott is making the Bounded Distrust argument that this doesn't much matter. That argument would need to apply equally to Donald Trump. And it seems like it does, in the sense that there are some statements he makes and I think he's probably giving me new true information, and other times he makes statements and I don't think that, and there's a kind of concreteness that's a large part of the distinction there.

And who is indeed *sometimes* bound to reality, and also often plays the game by exactly these rules. ‘Many people’ are saying X, you see, a lot of people, very good people. But not Trump, not directly, because *those are the rules of the game*. Similarly, when Cohen testified he noticed Trump being very careful with his word choices in private conversations, for similar (legal) reasons. Trump will also outright lie, of course, but he *is a politician and a real estate developer* so please don’t act quite so surprised and outraged. And he too is playing a game of this type and will choose when the price is too high and when it isn’t. The only difference is that he managed to ‘make a deal’ and thus pays lower prices. So he buys more and more brazen falsehoods, and occasionally he picks a falsehood that feels some combination of worthwhile and true to him and decides to double down on it.

Which is, as everybody knows, the rule for politicians. Who, with notably rare exceptions, will lie, to your face, all the time, about actual everything.

It’s worth noting that [the linked-to report from Wisconsin](#), also in WaPo, was better on many dimensions, not perfect but definitely coming from a world in which there is more focus on physical world modeling and less on narrative.

When I focus purely on the facts from this article that seem like sufficiently detailed non-trivial physical claims that they have content that we could potentially rely upon, *and edit to take out all the dripping contempt and hatred*, what’s left is this.

1. The [AP’s assessment](#) of the 2020 election uncovered 473 questionable ballots.
2. “He said a soon-to-come report from a source he would not disclose would support his case,” the AP reported Trump saying. Trump did respond with: “I just don’t think you should make a fool out of yourself by saying 400 votes.”
3. A total of 25.6 million cast ballots were cast in the states analyzed by the AP.
4. We have multiple state-level reviews conducted by Trump allies suggesting that the vote totals in [contested states](#) was legitimate. There has been no person who has stepped forward and admitted participation in any sort of scheme to throw the election and no discovery of rampant, coordinated fraud save for an effort to [cast ballots in Macomb County, Mich.](#), that constitutes most of AP’s total from that state — an effort that didn’t actual result in ballots being counted. And then there’s AP’s broad analysis of the vote in all six states that found only piecemeal problems.

Here’s an important thing *not* on the list:

It often takes a while for states and counties to adjudicate dubious ballots. It’s a lengthy process, matching cast votes with actual voters. But counties have a sense now of how often votes might have been cast illegally. In sum: fewer than 500.

Because that, you see, *is allowed to be functionally false*, and also actually is functionally false, conflating different numbers at *least* three times.

It’s conflating the ballots cast in the states analyzed by the AP – 25.6 million – with the combined number of ballots cast, which was about five times that number. Whereas the AP analyzed only a subset of those 25.6 million ballots. And it is then implicitly stating that there is zero chance that any ballot *not* viewed as suspicious by the AP could have been cast illegally. While the chance of any given ballot cleared by the AP having been cast illegally is very low, there are ways to do this that would not show up on the ballot itself, and that would not have been detected.

When you're willing to make this level of misstatement about the core question at issue, it makes it that much harder to know where you can still be credible.

Essentially what this is saying is:

1. The number I'm giving you comes from *somewhere*.
2. It doesn't have to be the thing you naturally think it is.
3. That number can represent a subset or otherwise be heavily misleading.

The window of what we are forced to treat as real keeps narrowing.

The original version of our Fact Statement #3 was, in fact, this:

Even if every one of those 473 cases was an actual example of fraud, it's out of a total of 25.6 million cast ballots.

Which implies that those 25.6 million ballots were all analyzed as part of the AP's work. They weren't. I had to realize this and back-edit to fix that.

Clicking through to the AP article provided clarity on many things, but still the whole thing boils down to whether or not you trust the Associated Press to do an investigation like this. I don't think it makes you a 'crazy conspiracy theorist' to think that the AP was not, via this method, going to detect all or even most potential forms of fraud that might have taken place.

If I imagine to myself that Omega (a hypothetical omniscient omnipotent always fully honest entity) told me the election was fraudulent somehow, and then I'm told that the AP report is about to come out, I notice I still expect the AP report not to find anything. If there was anything that they would have been forced to find that way, someone else would have already found it. The AP doesn't have to lie to simply not notice things, and given who they are I expect them to be very good at not noticing.

So all of this boils down to this:

1. Liberal sources continue to push narrative that there was no significant fraud.
2. Liberal sources continue to push narrative that all specific physical claims of significant fraud have been debunked.
3. Trump continues to promise that he'll come up with evidence Real Soon Now.
4. The evidence is not currently present, because otherwise Trump would say so.

That's true as far as it goes.

And you know what? Points three and four are actually *really super strong* evidence that no one has this kind of concrete evidence of fraud. It's the kind of very specific claim – that Trump is not saying X – that if false would be rapidly exposed as false because Trump would be very clear he was doubling down on X and this would be reported.

Thus, when Scott says this:

In order to avoid becoming a conspiracy theorist, the conservative would have to go through the same set of inferences as the FOX-watching liberal above: this is a terrible news source that often lies to me, but it would be surprising for it to lie *in this particular case in this particular way*.

I say no, absolutely not. The article in question is saying a mix of that which is, and that which is not, and a lot of that which is but is mostly designed to imply that which is narratively convenient without regard to whether or not it is true.

You can reasonably argue that there are *particular statements* within the article that one can be highly confident from context are being stated accurately. But one can accept those particular statements without it forcing one to accept that the election wasn't stolen. There's no logical incompatibility here.

Part of what's going on is that this 'conspiracy theorist' label is a threat being used, and you have to do things to avoid being labeled that way. In particular, you need to notice what everybody knows is a conspiracy theory right now, and avoid advocating for it. If that changes and something (for example, the lab leak hypothesis or UFOs) stops being considered a conspiracy theory, you can then switch your tune.

Things like this Washington Post article tell us nothing we don't already know. All of the work is relying on this line and similar logic:

The 2020 election got massive scrutiny from every major institution.

The core argument is that the absence of evidence is, in this context with this many people looking this hard to find something, and with the magnitude of the necessary efforts to pull this off and the resulting amount of evidence that would be available to be found, and the number of people who could potentially talk, very strong evidence of absence. That the amount of 'evidence of fraud' that was found is about the amount you'd expect to find if there was no significant fraud and this kind of effort, if anything it's less than that. It's surprising that, even with nothing to find, stuff more suspicious than this couldn't be found.

One could say that the liberal media would suppress such findings, and no doubt some parts of it would attempt to do so if such findings arose, but there are enough media sources on the other side that we need not worry much about such suppression happening without being noticed.

The liberal media *could and did* essentially use their One Time on Donald Trump in various ways, and paid the price in future credibility for doing so, but even with that it wouldn't have been enough to sell us a centrally fraudulent election in a way that couldn't be noticed.

All of the claims of fraud were even politely registered in advance as claims that would be made no matter what if the wrong side won, so they're actual zero evidence of anything except in their failure to be better substantiated. Whereas if there was big fraud, we would almost certainly know. And the reason for that is that the ones claiming fraud realized that the distrust of institutions was no longer sufficiently bounded to convince people not to believe such fraud claims, so there was no incentive not to make the claims regardless of the degree of fraud.

Combine that with a reasonable prior, and you get extremely high confidence of no fraud.

What you don't get is especially bounded distrust in the media sources involved.

As I was writing this Marginal Revolution linked to [this excellent post about why the USA is unlikely to face civil war](#). Among other things, it notices that various measurements of America's democracy were altered to make Trump look scary in

ways that don't make *any* sense. Then America's past was *retroactively* made worse to make it consistent with the ratings they gave for modern America to make Trump look maximally bad and also get in digs on the outgroup while they were at it. You can fairly say once again, blah blah blah, none of that is specific actual physical world falsifiable claims so of course all such things were pure political propaganda, physical world falsifiable claims are different. But this kind of thing is then cited as a 'source' to back up claims and sounds all scientific and tangible even though it's not, and is an example of something that wouldn't have happened twenty years ago (as, in they *went back and did it retroactively* because it was too fair back in the day) so it's an example of the war on memory in such matters and also the decay of the bounds of distrust. And also, these are 'experts' giving their opinions, so now 'experts' who aren't giving physical world falsifiable (in practice, not in theory) claims need to *also* be ignored by that standard.

Basically, I'm saying no, you *can't* evaluate any of this by saying 'look at all these experts' and 'look at all these *institutions*' without also using your brain to think about the situation, the counterfactuals and the likelihood ratios of various observations and applying something that approximates Bayes Rule.

I also am pretty sure that Scott did exactly the thing where you at least implicitly calculate a bunch of likelihood ratios of various observations and applied Bayes Rule and came to the same conclusion as everyone else who did this in good faith in this case.

Science™

Scott tells the story this story.

According to [this news site](#), some Swedish researchers were trying to gather crime statistics. They collated a bunch of things about different crimes and – without it being a particular focus of their study – one of the pieces of information was immigration status, and they found that immigrants were responsible for a disproportionately high amount of some crimes in Sweden.

The Swedish establishment brought scientific misconduct cases against the researchers (one of whom is himself "of immigrant background"). The first count was not asking permission to include ethnicity statistics in their research (even though the statistics were publicly accessible, apparently Swedish researchers have to get permission to use publicly accessible data). The second count was not being able to justify how their research would "reduce exclusion and improve integration."

It counts as 'scientific misconduct' for you to not be able to justify how your research would 'reduce exclusion and improve integration.'

Which is odd.

It means it is official policy that wrongfacts are being suppressed to avoid encouraging wrongthink and wrongpolicy.

It also means that we can no longer have a thing called 'scientific misconduct' that one can use to identify sources one cannot trust, since that now could refer to wrongfacts. If someone says 'that person is accused of scientific misconduct' I need to

be very careful to get the details before updating, and if I don't I'm effectively reinforcing these patterns of censorship.

But, Scott says, scientists have the decency to accuse them of misconduct *for failure to reduce exclusion*. This has the benefit of making it clear that this is an act of censorship and suppression rather than that the scientists did something else wrong, for anyone paying attention. If the claims were false, the scientists cracking down on wrongfacts would say the facts in question were wrong. By accusing someone of saying wrongfacts but *not saying the wrong facts are wrong*, you're essentially admitting the wrongfacts are right. So this gives you, in this model, something to go on.

I believe that *in some sense*, the academic establishment will work to cover up facts that go against their political leanings. But the experts in the field won't lie directly. They don't go on TV and say "The science has spoken, and there is strong evidence that immigrants in Sweden don't commit more violent crime than natives". They don't talk about the "strong scientific consensus against immigrant criminality". They occasionally try to punish people who bring this up, but they won't call them "science deniers".

Let me tell you a story, in three acts.

1. All masks don't work unless you're a health professional.
2. All masks work.
3. Cloth masks don't work.

At each stage of this story, scientists got on television to tout the current line. At each stage of this story, the 'science denier' style labels got used and contrary views were considered 'dangerous misinformation.'

Yes, we did learn new information to *some extent*, but mostly we knew the whole story from the beginning and it's still true now. Cloth masks are substantially better than nothing, better masks are much better. Also the super-masks like P100s (or the true fashion statements that work even better) are far better than N95s and you're basically never allowed to mention them or advocate for mass production. And yeah, we all knew this back in March of 2020, because it's simple physics.

I could also tell you a story about vaccines. Something like this:

1. Vaccines are being rushed.
2. Vaccines are great and even prevent all transmission and you're all set.
3. Vaccines are great but you still have to do all the other stuff and also you need a booster even if you're a kid unless you're in one of the places that's illegal. But only the one, definitely, that's all.

And that's entirely ignoring the side effect issue.

Once again, yes, you could say that the information available changed. On boosters, I'm somewhat sympathetic to that, and of course Omicron happened, but don't kid yourself. Motivations changed, so the story changed.

Then there's the lab leak hypothesis. And the *other* lab leak hypothesis.

Then there's social distancing and 'lockdowns' and protests where the scientists declared that social justice was a health issue and so the protests weren't dangerous.

Which are words that in other contexts have meaning.

Then there's the closing of the schools and remote learning and telling us masks and the other stuff isn't doing huge damage to children.

Then there's travel restrictions.

There's the WHO saying for quite a long time that Covid isn't airborne.

There are the claims early on of 'no community spread' while testing was being actively suppressed via the CDC requiring everyone to use only its tests when it knew they didn't work.

There's Fauci saying we'd get to herd immunity at one number, then saying that when we'd made enough progress on vaccination he felt free to increase the number a bit more, indicating he didn't care about what the real number was. And he wasn't alone.

And so on.

And in each case, the relevant 'expert' people who are wearing official 'trust the science' lapel pins explicitly lied, over and over again, using different stories, right to our f***ing faces. While arranging for anyone who disagrees with them to be kicked off of social media or otherwise labeled 'dangerous misinformation.' Then they lied and say they didn't change their story.

So when we say that scientists 'don't lie directly' we need to narrow that down a bit.

Can we say 'don't lie directly about specific actual physical world falsifiable claims?'

I mean, no. We can't. Because they did and they got caught.

There's still some amount of increasing costs to increasingly brazen misrepresentations. That's why, in the Swedish example, we don't see direct false statements to deny the truth of the claims made. The claims made are too clearly true according to the official statistics, so opening up yourself like that would only backfire. But that's a tactical decision, based on the tactical situation.

This is, as Scott says, a game with certain rules. But not very many.

If there is a published paper or even pre-print in one of many (but not all) jurisdictions, I mostly assume that it's not 'lying about specific actual physical world falsifiable-in-practice-if-false claims.'

Mostly. And that's it. That's *all* I will assume about the paper.

I will not assume it isn't p-hacked to hell, [that it has any hope of replication](#), that anything not explicitly mentioned was done correctly, that the abstract well-described the methodology or results, that their discussion of what it means is in good faith, or anything else, except where the context justifies it. I may choose to do things like [focus on the control variables](#) to avoid bias.

Outside of the context of an Official Scientific Statement of this type, even more caution is necessary, but *mostly* I still would say that if it's something that, if false, I could prove was false if I checked then the scientist will find a way to *not quite* say the false thing as such.

So yeah, anthropogenic global warming is real and all that, again *we know this for plenty of other good reasons*, but the reasoning we see here about why we can believe that? No.

And that suggests to me that the fact that there *is* a petition like that signed by climatologists on anthropogenic global warming suggests that this position is actually true. And that you can know that – even without being a climatologist yourself – through something sort of like “trusting experts”.

This is not the type of statement that we can assume scientists wouldn’t systematically lie about. Or at least, it’s exactly the type of statement scientists will be rewarded rather than punished for signing, regardless of its underlying truth value.

That’s mostly what the petition tells you. The petition tells you that scientists are being rewarded for stating the narrative that there is anthropogenic global warming. And they would presumably be severely punished for saying the opposite.

Both these statements are clearly true.

The petition does not tell you that these people *sincerely believe* anything, although in this case I am confident that they mostly or entirely do. It *definitely* does not tell you that these people’s sincere beliefs are right, or even well-justified, although *in this case* I believe that they are. This kind of petition simply does not do that at this time. Maybe we lived in such a world a while ago. If so, we live in such a world no longer.

But why am I *constantly putting in those reminders that I am not engaging in wrongthink*? Partly because I think the wrongthink is indeed wrong and I want to help people have accurate world maps. Partly to illustrate how ingrained in us it is that there is wrongthink and rightthink and which one this is, and that this petition thus isn’t providing much evidence. And partly, because I *really don’t want to be taken out of context* and accused of denying anthropogenic global warming and have that become a thing I have to deal with and potentially prevent me from saying other things or living my life. Or even have to answer the question three times in the comments. And while I don’t *think* I was in any danger of all that here, I can’t be sure, so better safe than sorry.

In my case, if I believed the local wrongthink, I would avoid lying by the strategy of being very very quiet on the whole topic because I wouldn’t want to cash in this type of One Time on this particular topic and risk this being a permanent talking point whenever my name comes up. Wouldn’t be Worth It.

Others are surely thinking along similar lines, except not everyone has the integrity and/or freedom to simply say nothing in such spots. In any case, no, the petition did not tell me anything I did not already know, nor do I expect it to convince anyone else to update either.

Then Scott goes on to say this.

(before you object that some different global-warming related claim is false, please consider whether the IPCC has said with certainty that it isn’t, or whether all climatologists have denounced the thing as false in so many words. If not, *that’s my whole point.*)

So it *sounds like* the standard is specifically that *the IPCC does not make statements that false things are definitely true*. Whereas if ‘some climatologists’ make such

claims, that's unsurprising. So when enough scientists of various types go around saying we are *literally all going to die* from this and manage to convince a large portion of an *entire generation* to think they are so doomed they will never get to grow old, we can't even treat that as evidence of anything, let alone call them out on that, because the IPCC hasn't specifically said so. I mean, [I checked](#) and they don't appear to have said anything remotely similar.

Yet I don't see them or any other 'experts' standing up to boldly tell everyone that yes we have much work to do but maybe we can all calm down a bit. And maybe we should avoid the overselling because it will cause people to think such 'experts' can't be trusted. Whereas I see other 'experts' adding fuel to this fire, presumably because they think that only by getting people into that level of panic can they get people to actually *do something*. A potentially noble motive to be sure, depending on details and execution, but not exactly the names you can trust.

Some people wonder how so many people could not Trust the Science™ in such matters. I don't wonder about that.

Nor do I think this is the reason Scott believes in AGW. Does Scott look like the type of person who says 'oh all these experts signed a statement so I'm going to believe this important fact about the world without checking?' No. No he does not. Scott is the type of person who actually looked at the evidence and evaluated what was going on for himself, because that's what Scott does and the only mystery is how he does so much of it so quickly. Even for me, and by not-Scott ordinary-human standards I do a lot of analysis very quickly.

Ivermectin One Last Time Oh Please God Let This Be The Last Time

Last year I explained why I [didn't believe ivermectin worked](#) for COVID. In a subsequent discussion with Alexandros Marinos, I think we agreed on something like:

1. If you just look at the headline results of ivermectin studies, it works.
2. If you just do a purely mechanical analysis of the ivermectin studies, eg the usual meta-analytic methods, it works.
3. If you try to apply things like human scrutiny and priors and intuition to the literature, this is obviously really subjective, but according to the experts who ought to be the best at doing this kind of thing, it doesn't work.
4. But experts are sometimes biased.
5. F@#k.

In the end, I stuck with my belief that ivermectin probably didn't work, and Alexandros stuck with his belief that it probably did. I stuck with the opinion that it's possible to extract non-zero useful information from the pronouncements of experts by knowing the rules of the lying-to-people game. There are times when experts and the establishment lie, but it's not all the time. FOX will sometimes present news in a biased or misleading way, but they won't make up news events that never happen. Experts will sometimes prevent studies they don't like from

happening, but they're much less likely to flatly assert a clear specific fact which isn't true.

I think some people are able to figure out these rules and feel comfortable with them, and other people can't and end up as conspiracy theorists.

A conspiracy theorist, officially now defined as anyone believing the Official Lying Guidelines are more flexible than *you* think they are (see: everyone driving slower than me is an idiot, anyone driving faster than me is a maniac).

Scientists engaging in systematic suppression of Ivermectin trials via various tactics? Well, of course. Scientists making *certain specific kinds* of false statements that *go against the 'rules'*? Conspiracy theory. Even though the rules keep loosening over time, and sometimes some things labeled 'conspiracy theory' turn out true, and also many things labeled 'conspiracy theory' don't actually even require a conspiracy, that's just a way of dismissing the claims.

Scott wrote a long post about Ivermectin. In that post, did Scott rely on 'experts' to evaluate the various papers? No, he most certainly did not. Scott *actually looked at the papers* and considered the evidence on each one and made decisions and then aggregated the data. And then, after all that, he took a step back, looked holistically at the situation, [found it best matched a hypothesis from Avi Bitterman](#) (worms!) and went with it, despite no 'experts' having endorsed it, and then a lot of people went 'oh yeah, that makes sense' and adopted the conclusion, which is how this works, is exactly how all of this works, that's Actual Science rather than Science™.

As in, yeah, step three above is *true*, the 'experts' definitely reach this conclusion. But also we looked at *exactly why* those experts got to that conclusion, and story checks out. Also Scott looked in detail himself and got a more interesting but fundamentally similar answer.

Yes, experts are sometimes biased, if you're being charitable, or 'engaged in an implicitly coordinated suppression of information in conflict with the current narrative' if you're being more realistic. Also, sometimes they're simply *wrong*, they have limited information to work with and limited cognition and lousy incentives and lives and this whole science thing is hard, yo. That's why Scott had to spend countless hours doing all that work for himself rather than 'Trusting the Science™.' Which looks a lot different than 'the experts wouldn't lie about this particular thing so of course Ivermectin doesn't work.'

I mean, the experts still haven't come around to the Vitamin D train, so 'the experts aren't impressed by the evidence' isn't exactly what I'd think of as a knock-down argument against non-risky Covid treatments.

Also, remember the rules that Scott mostly agrees upon. The scientists aren't allowed to say anything provably false, but they *are* allowed to suppress studies and other information they don't like by making isolated demands for rigor.

Which is *exactly* what Alexandros claims they are doing. I can confirm this more generally because I spent a bunch of time talking to him as well. Then, in Alexandros' model, having raised enough FUD (fear, uncertainty and doubt) around the studies in question, and using that to cast doubt on any that they couldn't do hit jobs on, they go and say 'no evidence' which is a standard accepted way to say that which is not, and that's that. You don't even have to tell the scientists explicitly to do that because they notice the narrative is that Ivermectin is outgroup-branded and doesn't work, and

that's that. In all my conversations with Alexandros, I can't remember him ever claiming any scientist outright lied *in the way Scott says they don't lie*. His story in no way requires that.

Which, again, is why Scott had to spend all that time looking himself to know for sure.

Once again, I agree with Scott on the bottom line. As far as I can tell, Ivermectin doesn't work.

But once again, I don't think Scott's stated algorithm is a good one, although once again I happily *don't think Scott is using his stated algorithm* in practice. I think he's *mostly* using mine, with the main difference being that I think he hasn't sufficiently adjusted for how much the goalposts have been moved.

The *real* disagreement between Scott and Alexandros here is exactly that. Alexandros thinks that scientists suppressed Ivermectin using arguments they would have been able to successfully make in exactly the same way whether or not Ivermectin worked. Thus, he claims that those arguments provide no evidence against Ivermectin, whereas there is other evidence that says Ivermectin works. Scott thinks that there are enough hints in the details and rigor of the arguments made that yes, they constitute real and strong evidence that Ivermectin does not work.

More likely, Scott noticed that the people pushing for Ivermectin were part of the Incorrect Anti-Narrative Contrarian Cluster who also push a bunch of other anti-narrative things that are not true, rather than part of the [Correct Contrarian Cluster](#) (CCC). There weren't people who otherwise were playing this whole game correctly but also happened to buy the evidence for Ivermectin. Whereas those who advocated for Ivermectin were reliably *also* saying vaccines were dangerous or ineffective, and other anti-Narrative claims that were a lot less plausible than Ivermectin, usually along with a bunch of various assorted obvious nonsense.

Which in turn meant that when one did look at the evidence, the cognitive algorithms that caused one to support Ivermectin were ones that also output a lot of obvious nonsense and were functioning to align and appeal to an audience with this uniform set of obvious nonsense beliefs, and when something in that group is investigated it turns out to be nonsense or in violation of one of the sacred Shiboileths, so it may be completely unfair and a potentially exploitable strategy but as a Bayesian when one sees something in that cluster *that doesn't violate an obvious sacred Shibboleth* it is safe to presume it is nonsense. And if it does violate a Shibboleth, then hey, it's violating a Shibboleth, so tread carefully.

One can (and whether one realizes it or not, one does to some extent) use it in the climate change example, noticing that full denial of climate change is very much part of the Incorrect Anti-Narrative Contrarian Cluster (ICC), while also noticing that moderate positions are conspicuously *not* in the ICC but rather in the CCC.

Of course, that's a level of attention paying and reasoning that's in many ways *harder* than doing the core work oneself, but it's also work that gets done in the background if you're doing a bunch of other work, so it's in some sense a free action once you've paid the associated costs.

One must of course be *very very careful* when using such reasoning, and make sure to verify if the questions involved are actually important. If you treat the CCC as true and/or the ICC as false than you are not following the algorithm capable of generating the CCC or rejecting the ICC. I mean, oh yes, this is all *very very* exploitable, as in it's

being exploited *constantly*. Often those trying to suppress true information will try to tar that information by saying that it is believed by the ICC. Although they are rather less polite and very much *do not* call it that.

But although all this did cause Scott to have a skeptical prior, Scott makes it clear that he came into his long analysis post not all that convinced. Hence the giant looking into it himself.

I also notice that Scott didn't choose any examples where the narrative in question is centrally lying to us, so it's hard to tell where he thinks the border is, until the final note about the harvest.

Glorious Harvests

Scott's next argument is that our Official Narrative Pronouncements can be thought of as similar to Soviet pronouncements, like so.

But also: some people are better at this skill than I am. Journalists and people in the upper echelons of politics have honed it so finely that they stop noticing it's a skill at all. In the Soviet Union, the government would say "We had a good harvest this year!" and everyone would notice they had said *good* rather than *glorious*, and correctly interpret the statement to mean that everyone would starve and the living would envy the dead.

Imagine a government that for five years in a row, predicts *good* harvests. Or, each year, they deny tax increases, but do admit there will be "revenue enhancements". Savvy people effortlessly understand what they mean, and prepare for bad harvests and high taxes. Clueless people prepare for good harvests and low taxes, lose everything when harvests are bad and taxes are high, and end up distrusting the government.

Then in the sixth year, the government says there will be a *glorious* harvest, and neither tax increases *nor* revenue enhancements. Savvy people breath a sigh of relief and prepare for a good year. Clueless people assume they're lying a sixth time. But to savvy people, the clueless people seem paranoid. The government has said everything is okay! Why are they still panicking?

The savvy people need to realize that the clueless people aren't *always* paranoid, just less experienced than they are at dealing with a hostile environment that lies to them all the time.

And the clueless people need to realize that the savvy people aren't *always* gullible, just more optimistic about their ability to extract signal from same.

I mean the clueless people aren't exactly wrong. The government is still lying to them in year six, in the sense that the harvest is unlikely to be what you or I would call 'glorious,' and they will doubtless find *some other* ways to screw the little guy that aren't taxes or revenue enhancements.

But if that's *all* it is, then the point is essentially correct. There are rules here, or rather there are incentives and habits. The people are responding to those incentives and habits.

That doesn't mean the 'savvy' position is reliable. Being savvy relies on being *unusually* savvy, and keeping track of how far things have moved. Every so often, the goalposts got moved, you think you know what 'good' or 'glorious' means, but you're using the *old* translation matrix, and now you're *wrong*, and often that's because people noticed the translation matrix people were using and wanted to control the output of that matrix.

Those rules are anti-inductive, in the sense that they *depend on the clueless remaining clueless*. If the clueless did not exist, then the statements stop serving their purpose, so they'd have to ramp up (or otherwise change) the translation system. At some point, the government cashes in a One Time to say 'glorious' instead of 'good,' the living still envy the dead, and now if the system keeps surviving 'glorious' means 'the living will envy the dead' and 'legendary' means we will get to put food on the table this year. Then at some point they cash that in too, and so on. In other less centralized contexts, this word creep is continuous rather than all at once.

Then at some point the translation system resets and you start again, with or without the system of power underlying it collapsing. One way for this to happen is if 'glorious' already means 'the living will envy the dead' and I say 'lousy' then that can't be intended to be translated normally, so I might actually honestly mean lousy without thinking the living will envy the dead, and so the baseline can reset.

But if you play this game, you *by construction* have to lose a large percentage of the people who will be confused what you're doing. It's designed to do that. One can't then look at the clueless and tell them to get a clue, because there's a fixed supply of clues.

If the system is distributed rather than centrally determined, and it's a bunch of people on social media running around labeling things as other things, then you see a gradual ramping up of everything over time as people adjust expectations and get wise to the game, or as the Narrative's forces win battles to expand their powers and then launch new attacks on the opposition. If I want to say something is glorious I have to be two steps ahead of whatever I view as the 'standard' description. Other similar dynamics exist in other places where words meanings can be changed or expanded over time, because those words serve purposes.

Bounds, Rules, Norms, Costs and Habits

Scott views bounded distrust as a game with rules and lines. There are some lines you mostly obey but sometimes cross at a price, and some lines you don't cross.

I'd modify that to say that there mostly *aren't* lines you simply do not cross. There are only lines that are *expensive to be caught* crossing when similar others are not *also caught* crossing them.

This is a variant of [having correlated debts](#), or losing money in the same way those around you lose money. You mostly only get punished for getting singled out as unusually bad. Thus, the more you are pushing *the same lies* as others and breaking *the same rules*, especially as part of The Narrative, you are effectively protected, and thus the price of breaking the rules is far lower.

When deciding what to do, various players will rely on some combination of bounds, rules, norms, costs and habits. Mostly, they'll do whatever they are in the habit of

doing, and those habits will adjust over time based on what is done, rather than thinking carefully about costs and benefits. This can also be thought of similarly as them following and over time changing the norms that are being locally and globally followed. They'll look at the costs and benefits of following or breaking what they think of as 'the rules' in various ways, mostly intuitively, and decide what to do about that in context.

Centrally, most official, news and 'expert' are looking to balance the opportunity to show their loyalty to and support the Narrative that they're getting behind, and the rewards for doing that, against the penalties that might be extracted if they are caught getting too far out of line and doing things that are out of line, and thus hammered down upon.

It is out of line to go too far and get caught, to be too far removed from the underlying physical reality in ways that can be observed or proven, and thus that weaken the Narrative and your reputation. You lose points for losing points, more than you lose points for anything else.

It is *also* out of line to *not go far enough*, and to adhere too well to what used to be 'the rules' rather than scoring sufficient Narrative points. One must stay on brand. This, too, is sticking one's neck out in a dangerous way.

The combination of these factors does often mean that there is effectively a *calibrated* response to any given situation. The details of what is said will be an intuitively but skillfully chosen balance of exactly what claims are made with exactly what level of specificity and rigor. Thus the chosen *details* of what is claimed and said actually can tell you quite a lot about the underlying physical world situation, *if* you can remain sufficiently well-calibrated in this and maintain the right translation matrix.

If you can do that, you can observe *exactly how much smackdown occurs and in exactly what way*, and know whether they're smacking down something true, something unclear or something false. The problem is that there's lots of inputs to that matrix, so without a lot of context you'll often get it wrong. And also the rules keep changing, so you need to keep your matrix up to date continuously.

Combining a variety of sources improves your results. Different sources, even with similar overall trustworthiness, will have *different* costs, both external and internal/intrinsic, and be pushing somewhat different Narratives. By observing the *differences* in their responses, you can learn a lot about what's going on by asking what would make all their responses make sense at once. Exactly who falls in line and in which ways, with what levels of weaseling, is no accident.

The principle that This is Not a Coincidence Because Nothing is Ever a Coincidence will serve you well here on the margin.

What Is the Current Translation Matrix?

I'm not going to justify this here, but seems only fair to tell where I am at. A full explanation would be beyond the scope of this (already very long) post, hence the incompleteness warning up front.

Here's mine for politicians:

They are on [what I call simulacra level 4](#), and they are moving symbols around without a direct connection to the underlying reality. Mostly, presume that politicians are incapable of means-ends reasoning or thinking strategically or engaging seriously with the physical world, and what comes out of their mouths is based on a vibe of what would be the thing one would say in a given situation, and nothing more.

Assume by default that they lie, all the time, about everything, including intentionally misstating basic verifiable facts, but that to model them as even thinking on those terms is mostly an error. Also assume that when they do say that which is not, if it is within the ability and the interests of the opposition to call them out on it then they will do so, and that the politician has intuitions that consider this and its consequences somewhat when deciding how brazenly to lie. While noting that in some situations, being called out on a lie is good for you, because it draws attention to the proper things and shifts focus the way you want.

Information about what type of vibe a politician is looking to give off is useful in terms of figuring out what vibe they are looking to give off, which can change when circumstances change. Explicit promises carry non-zero weight to the extent that someone would be mad at them for breaking those promises and that this would have felt consequences that can impact their intuitions, or other ways in which it directly constrains their behaviors.

Also assume that they will act as if they care about blame on about a two week time horizon, so the consequences of things being proven false mostly have to back-chain in time to punish them within two weeks, or no one will care.

And that's it.

For traditional news sources like the Washington Post, CNN or FOX:

Assume until proven otherwise that they are engaging primarily in simulacra level 3 behavior, pushing the relevant Narrative and playing to and showing their loyalty to their side of the dialectic to the extent possible. Thus, subject to the constraints they are under, assume they are giving the optimal available-to-them arguments-as-soldiers (also rhetoric-as-soldiers) version of whatever thing they are offering, and calibrate based on that.

Those constraints are a very narrow form of technically correct, the best kind of correct. Or rather, a very narrow form of *not* technically *incorrect*, with something that could be plausibly held up as some sort of justification, although that justification in turn need not be verified or accurate. So you can often have a circular information cascade with no actual evidence.

Basically, if a statement or other claim is:

1. A specific falsifiable claim about the physical world.
2. Could, if false, *in actual practice*, be falsified in a way that would 'count.'

Then it has to *technically* be laid out in a not false way, for example by saying that 'source Y (or an unnamed source) said that X' instead of X. The Marx/Lincoln story is an excellent example of exactly where this line is. Assume that like that story, everything will go *exactly up* to that line to the extent it is useful for them to do so, but not over it. Then, based on what content is included, you know they didn't have any better options, and you can back-chain to understand the situation.

Like politicians, they mostly also care about blame on a two-week time horizon, so there needs to be a way for the anticipated consequences of crossing lines and breaking rules to back-chain and be visible within two weeks, or they'll mostly get ignored.

Assume that they are constantly saying things similar to 'wet ground causes rain' when they want to be against wet ground, and also framing everything with maximum prejudice. Everything given or available to them will be twisted to inflict maximum Narrative (and get maximum clicks otherwise) wherever possible, and analyze output on that basis. Assume that they outright lied to their sources about what the story was about, or what information would be included, or anything else, if they found this to be useful and worth more than not burning their source. Also remember that if you are about to be a source.

Basically, yes, there is a *teeny tiny sense* in which they will not outright lie, in the sense that there is a Fact Checker of some kind who has to be satisfied before they can hit publish, but assume it is the *smallest sense possible* while still containing at least some constraint on their behavior.

Remember that any given 'source' can, for example, be a politician.

Remember that if the source is an 'expert' that means *exactly nothing*.

Also assume that headlines have (almost) zero constraints on them, are written by someone who really, really doesn't care about accuracy, and are free to not only be false but to *directly contradict the story that follows*, and that they often will do exactly that.

If information is *absent*, that *only* means that such information would have been unhelpful and they don't think it would be *too embarrassing* to simply ignore it, for which the bar is very high. They are under *zero obligation* to say anything they don't feel like saying, no matter how relevant.

If there's an editorial, there are no rules.

If it's in *any* way subjective, there are no rules.

Words mean whatever the Narrative decided they mean this week.

And that's it.

(I will note that in my experience, Bloomberg in particular does not do this, and can be trusted substantially more. There likely are also others like that, but this should be your default.)

For 'scientists' and 'experts':

If you want to find a 'scientist' or 'expert' to say any given thing, you can.

If you have some claim that fits the Narrative, then unless it is a full strict-false-and-one-could-prove-it violation, you can get lots of experts/scientists to sign off on it. So all you're learning is that this is part of the Narrative and isn't definitely false.

You *can* look at the details of the dissent and the details of what is in the petition or official Narrative statement, and exactly who conspicuously did/said or didn't say/do

what and exactly what weaseling is there, and extract useful information from that, because they're maximizing for Narrative value without going over the strict-false line.

Mostly any *given* expert will have *slightly* more constraints on than that, and will follow something similar to the news code, and will also have some amount of internal pressure that causes the vigor of endorsement to be somewhat proportional to the accuracy of the statement, but it's also proportional to the magnitude of the Narrative pressure being applied, so one must be cautious.

The more technical the talking gets, the more you can trust it (to the extent you can understand it), there's still some amount of dignity constraining behaviors in these ways in some places, but in other places it is mostly or entirely gone.

Also understand that the systems and rules are set up at this point to allow for very strong suppression of dissent, and creation of the illusion of consensus, through the use of social pressures and isolated demands for rigor and other such tactics, without need to resort to sharp falsifiable statements. Often the tactics and justifications involved in such moves are obvious nonsense when viewed by ordinary humans, but that is well within bounds, and *failing* to use such tactics is often *not* within bounds.

Expert consensus that is falsifiable-in-practice-in-a-punishing-way can still largely be trusted.

Expert consensus that is not that, not so much. Not as such. Not anymore. But you *can* sometimes notice that the consensus is *unexpectedly robust* versus what you'd expect if it wasn't trustworthy. You can also use your own models to verify that what the experts are saying is reasonable, combined with other secondary sources doing the same thing, and combined with *individual* experts you have reason to trust.

You should *definitely* expect the experts in any given field to greatly exaggerate the importance of the field at every turn, and to warn of the dire consequences of its neglect and our failure to Do Something, and for there to be real consensus on that for obvious reasons, except with less shame or restraint than in the past.

And, again, that's it.

There are other sources, specific sources, where the translation matrix is less extreme, and I of course do my best to draw as much as possible from such sources. There's still almost always a long ways to go before getting to the level of trust that would be ideal, but there are many levels.

So What Do We Do Now?

We decide how much time and effort we want to spend maintaining our calibration and translation matrix, and for which sources.

Maintaining a high-quality translation matrix of your own is a lot of work. That work isn't obviously worth it for you to do. There are three basic approaches here.

One is to basically *stop caring so much about the news*. This is a good strategy for many, and in most times. Before Covid, especially before Trump and when not doing any trading that relied on knowing what was going on, I was mostly implementing it. One can live the good life without caring about such matters. In fact, not caring often

makes it easier. Thus, you don't know what you can trust. But as long as you also *don't care*, it's fine.

You know what's going on hyper-locally, with your friends and family and work and neighborhood, and that's it. For most of history, that was enough.

This isn't as easy as staying away from newspapers and other official news sources. You *also* have to deal with the constant stream of news-bringing on social media, and in real life from coworkers, friends and family, and so on. You might want to be done with the news, but the news isn't voluntarily done with you.

You'll need to train yourself that when you see a post about today's terrible news, you ask yourself only one question. Will this directly impact the local physical world in ways that alter my life, thus forcing me to care? Or not? If not, move on. If it's political advocacy, or someone being wrong on the internet, definitely move on. Offline, you'll need to follow similar procedures, which will require smiling and nodding.

You'll also need to filter your incoming sources of non-news to filter out those who bring you too much news that isn't directly relevant to your life, and especially those who bring you political advocacy. This leads to some tough choices, as there are sources that have a combination of worthwhile things and exactly what you want to avoid. They're mostly going to have to go.

A second option is to keep very careful track of the physical world conditions, do lots of your own work and not need to rely on secondary sources like newspapers. I assure you that mostly this is a lot of work and you only want to do this in carefully selected sub-realms. It's taking the local approach and extending it to some non-local things, but it's difficult and it's time intensive, and mostly only makes sense if your conclusions are in turn going to be relied on by others. Also, it often needs to *complement* keeping up your translation matrix rather than *substituting* for it, as I can attest from experience.

The other option is division of labor and outsourcing.

If you can find a sufficiently trustworthy secondary source that analyzes the information for you, then you don't need to worry about the trust level of their sources. That's *their* problem.

Or to put it another way, you don't have to have a *fully general* translation matrix. You only need to have a translation matrix *for sources you want to get information from*. You get to choose your portfolio of sources.

That can be as simple as your spouse or a good friend that you know you can trust. There is of course a risk of telephone problems if there are too many 'links in the chain' but such costs are often acceptable. Using a personal source has the extra advantage that they can filter for you because they have a good idea what is relevant to your interests.

It can also aggregate various community sources. There's the obvious danger of information cascades here as there is elsewhere, as the upstream sources are still what they are, but it does provide *some* amount of protection.

You can also choose anything from one or more bloggers to a set of Twitter accounts to a newspaper, radio show or TV program you find to be unusually trustworthy. Or combine any or all of these and other sources.

I sometimes hear that someone has decided to outsource their Covid perspectives to me and my posts in this way. The posts are designed to allow you to think for yourself and reach your own conclusions, but also to save you the work of needing to maintain a detailed translation matrix while doing so, especially since I hope that the correct matrix for DWATV itself is very close to the identity matrix, except for the need to ‘translate into one’s own language’ since my way of framing and thinking about things has quirks and likely doesn’t exactly match yours. But that’s ideally about understanding rather than trust.

I have compiled a lot of sources over the years that I trust to be rather high up on a ‘pyramid of trust,’ meme version not currently ready for publication. This includes most (but not quite all) of my friends, since I value such trustworthiness and careful speaking highly, but even within that set there’s clear distinctions of how careful one needs to be with each source in various ways.

Everyone I list on my links and blogroll qualifies as someone I am mostly willing to trust. If they didn’t count as that, I wouldn’t list them.

That doesn’t mean I fully trust their *judgment*, or that it’s all created equal, but there’s a sense in which I can relax when engaging with such sources. There’s also, of course, a sense in which I *can’t* relax even when dealing with most of those sources, to varying degrees. I wish that were not so, but better to accept it than to pretend it’s not true.

The best sources, at least for my purposes, do an excellent job of being transparent about how trustworthy they are being in any given situation. Scott Alexander, as a prime example, is very good at this.

That’s the landscape on a personal and practical level.

Mostly I recommend, for keeping general tabs on the world, collecting a list of sources you’ve decided you can trust in certain ways, and then mostly trusting them in those ways while keeping an eye out in case things have changed. Then supplementing that with one’s own investigations when it matters to you.

For keeping tabs on your own local world, there are no shortcuts. You’ll have to do the work yourself.

But what about the global problem as a global problem? Sure, politicians have mostly always lied their pants on fire, but what to collectively do about this epic burning of the more general epistemic commons?

There are no easy answers there.

My blog is in part an attempt at an answer. This seems very much like a Be The Change You Want to See in the World situation. Thus, one can begin by striving to:

1. Being a trustworthy source of information to the extent you can manage. This includes not silently dropping information whose implications you dislike.
2. That means being clear on how and why you believe what you believe, and how high your confidence is in it.
3. Explicit probabilities are great when appropriate.
4. As is holding yourself accountable when you’re wrong.
5. Not rewarding untrustworthy sources, including with undue attention. When appropriate, make it clear in what ways they cannot be trusted, but mostly don’t

give them the oxygen of attention that they thrive on.

6. Rewarding trustworthy sources, including with attention, spread the word.
7. Focus on the physical reality, de-emphasize all versions of the Narrative. Look to figure out the gears underlying all this, and create common knowledge.
8. Make it clear you are doing this, to provide reason to follow suit.

This doesn't have to be about having a blog, or even a social media account, or the internet, or any kind of information projection at all. It's about how people at all levels interact, in the world, with people.

Note for Commenters: The no-politics rules are importantly *not* suspended here. Some amount of interaction with politics will be necessary. But beyond a clear emphasis on physical-world simulacra-level-1 considerations, advocacy of positions and partisan bickering remain right out. I stand by willing to use the delete button and potentially the ban hammer if necessary, while remaining hopeful they will not be necessary.

The metaphor you want is "color blindness," not "blind spot."

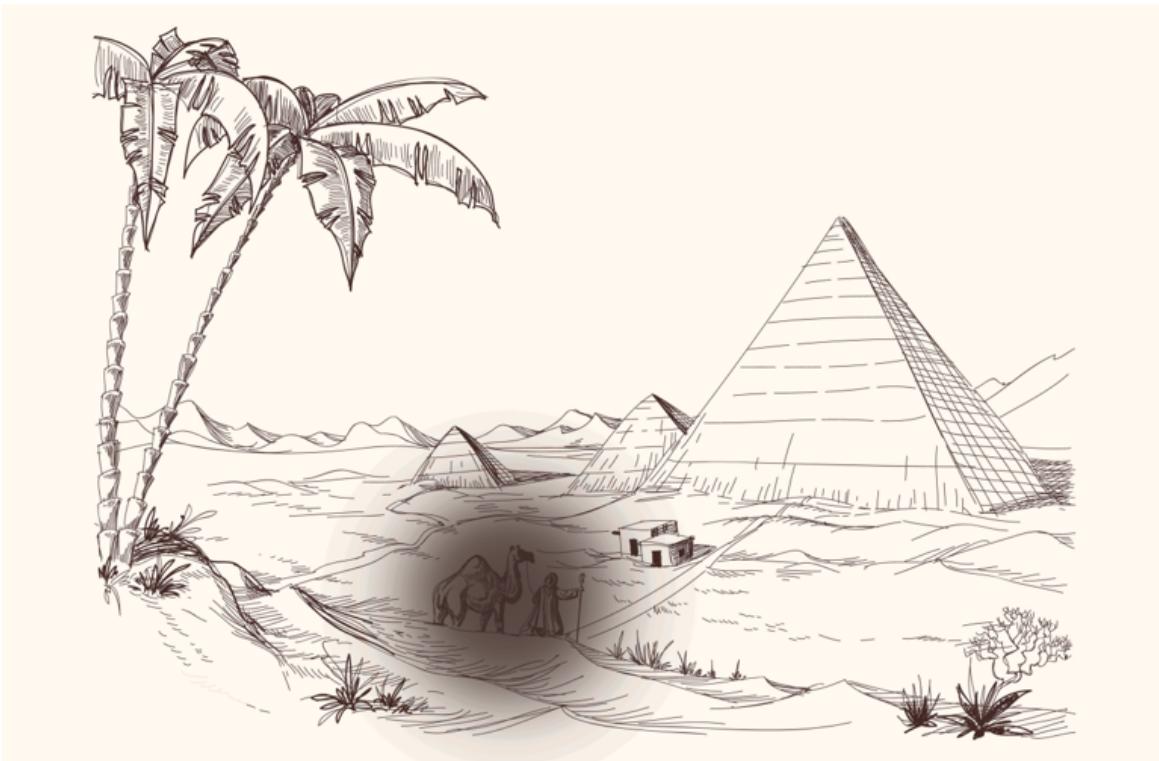
It's genuinely possible that the title is all you need; I was tempted to write nothing further. Feel free to take ten seconds and see if it's already clicked. If not, read on.

When something is in your literal blind spot, it's invisible to you, but your brain stitches together everything else around it to make you *think* that you're seeing a complete picture.

Reality:



Blind spot:



Perception:



We often use "blind spot" as a metaphor to gesture toward things we are unaware of, while also being at least somewhat unaware of *our unawareness*—we know that something fishy is going on, but we can't quite get our eyes on it.

e.g. "I think I might have a blind spot when it comes to status dynamics."

The thing about status dynamics, though, is that they aren't in *one spot*. There isn't a whole world that is being fully and accurately perceived, except for one blank space that's being glossed over.

Instead, what's usually going on (at least in my experience) is that the person can see *everything*, but there's some crucial component of the picture that they are unable to process or comprehend.

What this looks like, in practice, is an inability to distinguish two things which are very, very different, à la red-green color blindness:



"Look," says Alexis. "Look at the beautiful contrast."

Blake hesitates. "...you mean between the trees and the sky?"

Being (metaphorically) color-blind to something can be deeply frustrating. You keep pushing the X button, and *very different things* keep happening.

e.g. you are learning to play a Formula 1 racing simulator, and it *feels like* you did exactly the same thing on exactly the same curve both times, but one time you spun out and crashed and the other time you smoothly navigated right through.

e.g. you are repeating back to the native French speaker exactly the sounds she said to you, and sometimes getting nods and smiles and other times getting sympathetic winces.

e.g. you asked your romantic partner what you should have said, instead, to avoid this huge disagreement, and the words they wished you'd said are *literally equivalent in meaning, that's the exact same sentiment, what the hell is going on, here?*

It's straightforwardly analogous to being literally color-blind, where you might try on one t-shirt, and hear snickers and giggles and see people trying not to laugh, but then you swap it out for a t-shirt that is almost exactly the same shade, you can barely even tell them apart, and suddenly everybody's all encouragement and compliments.

They're the exact same shirt! They're practically indistinguishable!

Yes—to you. But other people make distinctions you do not. Whether because they've got slightly different physical or mental structures, or because they've spent a lot of time focusing on a domain and developed substantial sensitivity to it, or because they came from a subculture where those distinctions were obvious and omnipresent.

(And you make distinctions that are invisible to them, most likely! That seem to them capricious and arbitrary, if not entirely made up!)

Trigger: Notice that someone is being weirdly intense about a meaningless distinction.

Action: Ask "if they were perceiving some facet of reality that I am insensitive to (rather than just making mountains out of molehills), what might it be?"

Trigger: Notice that someone (including you) just used the term "blind spot."

Action: Ask what the thing allegedly "in" the "blind spot" is, and whether it is local, or distributed/omnipresent.

Trigger: Notice that someone is continually ignoring your gestures toward an Extremely Important Thing™, or that they keep failing to successfully extrapolate even though you've laid out a bunch of individual examples and the trend should be obvious by now.

Action: Imagine that they are *color-blind* to this property, and consider what other proxies or algorithms you might hand them, that are *not* dependent on directly perceiving the Extremely Important Thing™.

For calibration: mostly-replacing my mental category of "blind spot" with "color-blind" has been one of the ten most useful shifts of my last three years, by virtue of a) giving me a much better starting point for navigating inferential gaps, and b) somehow causing me to be more empathetic? ...I suppose because if the problem is a blind spot, then if they just shift their focus a little, they should *see the thing, dammit*, whereas if I model them as simply lacking the ability to perceive (even if only because they haven't practiced enough in this domain yet) it suddenly seems much less their-fault if merely shifting their focus doesn't fix it.

YMMV, but at the risk of being too clever by half: there really *is* a useful distinction, here.

Monks of Magnitude

Sometimes I encounter a concept and it immediately embeds itself [in my culture](#), such that it feels like I always knew it. This is a short description of such a concept; it was a near-perfect match for a way I already thought about things, and has since become a useful handle that I find myself explaining to others (so that I can subsequently reference it) roughly once a month.

There is a certain story, which I will not name here in order to reduce the spoiler-y nature of the following description.

In that story, there is a monastery, and the monastery is divided into tiers, or levels.

(I may not be precisely representing the story, here, but rather how the thing in the story ended up being recorded by my brain.)

Some monks are 1-day monks. They come out from their seclusion once a day, and mingle with the regular people, and share their insights, and make new observations, and then retreat back to their private spaces to muse and meditate.

Some monks are 10-day monks. They are much like the 1-day monks, except they come out only every 10 days.

The 10-day monks tend to think longer thoughts, and wrestle with subtler or more complex problems, than the 1-day monks. This is treated by the culture of the monastery as natural and correct. *Of course* the 10-day monks address a different set of problems; if 10-day monks and 1-day monks were good for the same purposes, the two different Orders wouldn't need to exist.

There are also 100-day monks, who come out only a few times per year.

There are also 1,000-day monks, who come out only every few years.

There are also 10,000-day monks, who come out only every thirty years or so.

There are also 100,000-day monks (the structure of this society has led them to be better at solving problems overall, which has allowed for some advances in longevity tech).

"Come out" may be a bit of a misnomer; in fact, it is the case that the most valuable insights of a given Order tend to be fully comprehensible only to monks of one, *mmmmaaaaybe* two Orders below them. So the 100,000-day monks, when they report in, mostly speak only to the 10,000-day monks, who are responsible for distilling and transferring relevant insights to the 1,000-day monks, who are responsible for distilling and transferring relevant insights to the 100-day monks, etc.

(It's also the case, as some readers have pointed out in the comments below, that certain problems cannot be solved by isolated thought alone, and require feedback loops or regular contact with the territory. For monks working on such problems, it is less that they sequester themselves completely for thousands of days at a time and more that, during those thousands of days, none can make demands of them.)

Duncan-culture works this way.

(By "Duncan-culture," I mean a culture composed entirely of Duncans; a culture made up of people who, whatever their other differences, take for granted everything that I, Duncan

Sabien, find intuitively obvious and believe I could convey to a ten-year-old version of me in a few hours' time. This society lives on a large island a few hours' sailing off the coast of dath ilan.)

If there were indeed 1,000 literal Duncan-copies available, to found a monastery or any other endeavor, they would immediately stratify themselves into 1, 10, 100, and 1,000-day groups at the very least, and probably there would be nonzero 10,000-day Duncans as well.

The key here is that each of these strata focuses on a set of largely non-overlapping issues, with largely non-overlapping assumptions.

To a 1-day or 10-day monk, questions like "maybe this is all a simulation, though" are almost entirely meaningless. They are fun to ponder at parties, but they aren't relevant to the actual working-out-of-how-things-work. 1-day and 10-day monks take reality as it seems to exist as a given, and are working *within* it to optimize for what seems good and useful.

But (of course!) we want *some* people working on 1,000 and 10,000-day problems! We don't want to *miss* the fact that this is all just a simulation, if it is in fact a simulation. And we don't want to be blind to the implications and ramifications of that fact, and fail to take appropriate action.

So *some* Duncans are off in the ivory tower, questioning the very fabric of reality itself, because *what if?*

And other Duncans are in between, taking different subsets of things for granted, while questioning others.

And it's fairly important that the 1,000 and 10,000-day monks *not be distracted* by such trivial concerns and questions like "how do we navigate continued cooperation within small groups after people have messy romantic breakups?"

(Or, well, most of them, anyway. Some small number of 10,000-day monks may in fact pay very close attention to exactly those dynamics, because those dynamics might contain Secret Subtle Clues As To How Things Really Work. There is no restriction, aesthetic or social or otherwise, on a higher-Order monk playing around with lower-order concepts to the extent that they find them useful or intriguing or refreshing or what-have-you.)

But for the most part, the 1,000 and 10,000-day monks are simply ... given what they claim to need. Their food and lodging is provided for; requests for companionship or certain odd materials are simply granted. The *assumption* is that most 1,000 and 10,000-day monks will produce nothing of measurable material value (especially not value comprehensible to a 1-day monk); the society as a whole has decided that it is nevertheless Extremely Well Worth It to fund *all* such monks, in perpetuity, for the once-in-several-lifetimes breakthroughs that *only* come from people who are willing to dive deeply into the terrifying Unknown.

Meanwhile, the 1, 10, and 100-day monks are busy improving the functioning of society, exploiting the current paradigm (rather than exploring in search of the next one). It is their labors which produce surplus and bounty and which, in a sense, "fund" the rest of the Orders.

(These distinctions are not clear-cut. The boundaries are fuzzy. This is fine; the monastery is sensible. Overall, though, the higher your Order, the less accountable you are to the bean-counters. Our current culture does something similar, though more clumsily, via e.g. tenured positions at universities.)

The reason Duncan-culture works this way is that it seems to be healthy, and sane. A culture with such a monastery, whose insights had repeatedly proven to revolutionize society, resulting in inventions like consistent judicial policy and microwave ovens and international

peace treaties and the general theory of relativity, is one that has *practiced* taking seriously ideas it does not fully comprehend. It's a culture that expects to sometimes be told "you do not understand why this is important, but it is." It's a culture that handles delegation via a chain of trust, and which e.g. "believes the science" in a way that does not devolve into mere tribal signaling whereby n95 face masks become a two-way shibboleth.

It's the *kind* of culture that e.g. would not fail to see global warming or existential risk from artificial intelligence coming, and would not fail to send the message to its elementary schools and universities "hey, we should start moving promising people into place to solve these problems" years or decades in advance of the deadline.

(There are other kinds of cultures that also avoid these failure modes, but they have other drawbacks.)

It's also the kind of culture that ... effortlessly navigates disagreement about what's important? You don't get criticisms of "ivory-tower nonsense" or "tunnel-visioned mundanity." People in such a culture understand, on a deep and intuitive level, that some problems are 1000-day problems, and other problems are 1-day problems, and *both are important*, and *both are important in very different ways*.



wtf is a conclusion paragraph. just stop reading

11:48 AM · Mar 3, 2021



540.1K Reply Copy link

[Read 734 replies](#)

(Just kidding, but in fact I don't have much of a tying-this-up-in-a-neat-narrative-bow conclusion. I think the concept is useful, and I think at this point you get it. My only parting recommendations are these: first, try categorizing the problems that catch your attention, and see if you tend to feel more-at-home in a particular Order. Second, try looking at various LW posts, and various prolific LW authors, and asking the question "if LW were such a monastery, which Order would this person belong to?" It makes the sometimes-disorienting diversity of LW content suddenly make a lot more sense, at least to me.)

(EDIT: Oh, a third one: "Are we mistakenly judging a 1,000-day monk by standards that only make sense for 10-day monks, or vice-versa?")

Harms and possibilities of schooling

To explore better possibilities for nurturing new minds, and to care about the problem in the first place, it helps to remember what's wrong with what we do to new minds. John Taylor Gatto speaks about this from experience and insight: [Seven Lessons Taught in School, 1991](#)

(If you're going to read the following, at least read the seven lessons (part I) of that essay.)

Here's another list of harms caused by schooling.

1. You aren't a mind, and don't bother trying to behave like one.

Children naturally attend to things until they're done with them:

From Maria Montessori, My System of Education, 1915: [IPFS pdf link](#)

A little girl, about three years of age, was deeply absorbed in the work of placing wooden blocks and cylinders in a frame for that purpose. The expression of her face was that of such intense attention, that it was almost a revelation to me. Never before had I seen a child look with such "fixedness" upon an object, and my conviction about the instability of attention which goes incessantly from one thing to another, a fact which is so characteristic in little children, made the phenomenon the more remarkable to me.

I watched the child without interrupting her, and counted how many times she would do her work over and over. It seemed that she was never going to stop. As I saw that it would take a very long time, I took the little armchair on which she was sitting and placed child and chair on the big table. Hastily she put the frame across the chair, gathered blocks and cylinders in her lap, and continued her work undisturbed. I invited the other children to sing, but the little girl went on with her work and continued even after the singing had ceased. I counted forty-four different exercises which she made, and when she finally stopped, and did so absolutely independently from an exterior cause that could disturb her, she looked around with an expression of great satisfaction, as if she were awakening from a deep and restful sleep.

The impression I received from the observation was that of a discovery. The same phenomenon became very common among those children, and it was noticed in every school in every country where my system was introduced; therefore it can be considered as a constant reaction which takes place in connection with certain exterior conditions that can be well established. Each time a similar "polarization" of the attention occurred, the child began to transmute itself completely; it became calmer, more expressive, more intelligent, and evidenced extraordinary interior qualities, which recalled the phenomena of the highest mentality. When the phenomenon of polarization of the attention had occurred, all that was confused and drifting in the conscience of the child seemed to assume a form, the marvelous characters of which were reproduced in each individual.

School usually steamrolls this process. The teacher has dominion over your attention; you have to either listen to what the teacher is saying, or work on the worksheets. Your location and (nominal) topic switches every hour. Even at home you do worksheets and projects. Your attention isn't you or yours; you're at war with your attention, it says no but your teacher--and implicitly your parents and college admissions and society, and "you"--insist.

Not only are you not the authority on what is worth it for you to spend minutes, days, or years on; if you are even consulted, it's superficial ("Which current event to do a report on? There are many options.") and maybe only for Potemkin village purposes ("We offer many electives."). You have to be a receptacle for what the teacher has brought to give you, because you don't have a perspective, you aren't an organized/organizing entity, and therefore can't be trusted to judge what is worthwhile. In the teacher-student relationship, someone has to defer to the other, and there's nothing in the student that the teacher could defer to. (I recall my elementary school art teacher looking at my painting, and then *taking the brush out of my hand and painting new stuff on the canvas over what I'd already put there*. In another class, I said "Wait, [...]" because I was confused about something, and the teacher interrupted my question to scold me for telling *zer* what to do.) Simply, the time in which you'd have created your mind is taken from you.

This creates learned helplessness about being absorbed in anything. It's like if you are trying to program, or write, but at arbitrary moments, someone Harrison Bergerons you until you forget what you were thinking. You learn not to get all worked up (absorbed in something, arranged ephemerally but suitably for the matter at hand, like a standing wave or a rough-and-ready scaffolding), because you'll just have to drop it in the middle anyway. You minimize the deepness and recursiveness of your questions, the length of your strides into the woods; the deeper the question, the more at risk you are of building towers of mental context and pumping neuroplasticity-juice into the relevant areas of your mind, and then having that plasticity act as random brain damage rather than successful reprogramming, when [the task that the plasticity was aimed at taking a compressive snapshot of how to perform/complete] is interrupted.

The self-organization of the child's mind is blocked at every turn, and the results are similar to trying to gestate a fetus in a small box.

[Always smile. Refrain from looking out of the window.](#)

2. The world isn't yours.

2.1. We own Space, and decide where you are.

Being forced to stay in a room with restricted range of motion (stay in your seat and stop fidgeting) is captivity, and captivity is harmful. Nuff said, one might have hoped.

A common episode: a student asks "Can I go to the bathroom?", and the teacher responds "You CAN, but you MAY not." or "You CAN, did you mean 'May I?'?". It's not just a stupid joke (what other stupid jokes stick out as vividly in memory as this one?), it's rubbing in your face that your range of motion is restricted by authority as if there

were a concrete wall instead of a door, but you're not allowed to bring to the teacher's attention that they are restraining you so severely. Wow, what a doofus you are to think that Can and May are the same, how could you possibly have gotten those confused, I wonder?

From [The Quiet Rooms, Richards, Cohen, Chavis, 2019](#):

TL;DR: Many thousands of times a year, kids in Illinois are locked up in a small room alone, sometimes for hours.

In Illinois, it's legal for school employees to seclude students in a separate space — to put them in "isolated timeout" — if the students pose a safety threat to themselves or others. Yet every school day, workers isolate children for reasons that violate the law[...] [snip]

For this investigation, ProPublica Illinois and the Tribune obtained and analyzed thousands of detailed records that state law requires schools to create whenever they use seclusion. The resulting database documents more than 20,000 incidents from the 2017-18 school year and through early December 2018. [snip]

"Please, please, please open the door. Please, I'll be good. Open the door and I'll be quiet."

"I'd rather die. You're torturing me."

Also, in Connecticut: ['Scream Rooms': Punishing Disabled Students in Isolation, Emily Richmond, 2012](#). Presumably in other states as well.

[Autistic 11 year old locked in a small room alone for hours in the UK](#): "He had always loved school... but by the end of October in fifth class he hated it. I was dragging him into school every day."

No kidding.

It sounds not uncommon in the UK: [Consequence Rooms](#). "Then he got 22 hours in an isolation booth in one week and he was just an absolute mess. He came out at the end of the day and he didn't look well. His legs were shaking and he could hardly string a sentence together. He looked completely done in."

Most schools aren't like that though, right? Well, the extremes, the ones that get reported, tell you about the hidden distribution and the attitudes that produced it. We own space, and decide where you ca... MAY go.

From Children's Games in Street and Playground, Opie and Opie, 1969: [IPFS pdf link](#)

TL;DR: Kids in confinement are more violent and cruel than kids not in confinement.

The places specially made for children's play are also the places where children can most easily be watched playing: the asphalt expanses of school playgrounds, the cage-like enclosures filled with junk by a local authority, the corners of recreation grounds stocked with swings and slides. In a playground children are, or are not, allowed to make chalk diagrams on the ground for hopscotch, to bounce balls against a wall, to bring marbles or skipping ropes, to play 'Conkers', 'Split the Kipper', 'Hi Jimmy Knacker'. Children of different ages may or may not be kept apart; boys may or may not be separated from girls. And according to the

closeness of the supervision they organize gangs, carry out vendettas, place people in Coventry, gamble, bribe, blackmail, squabble, bully, and fight. The real nature of young boys has long been apparent to us, or so it has seemed. We have only to travel in a crowded school bus to be conscious of their belligerency, the extraordinary way they have of assailing each other, verbally and physically, each child feeling—perhaps with reason—that it is necessary to keep his end up against the rest. We know from accounts of previous generations with what good reason the great boarding schools, and other schools following, limited boys' free time, and made supervised games a compulsory part of the curriculum. As Sydney Smith wrote in 1810, it had become an 'immemorial custom' in the public schools that every boy should be alternately tyrant and slave. [snip for length; more descriptions of abuse by kids in school]

[snip] [...] leading us [educators] to believe that a Lord of the Flies mentality is inherent in the young[...] [snip]

Thus recent extensive studies of apes and monkeys have shown, perhaps not unexpectedly, that animal behaviour in captivity is not the same as in the wild. In the natural habitat the welfare of the troop as a whole is paramount, the authority of the experienced animal is accepted, the idiosyncrasies of members of the troop are respected. But when the same species is confined and overcrowded the toughest and least-sensitive animal comes to the top, a pecking order develops, bullying and debauchery become common, and each creature when abused takes his revenge on the creature next weakest to himself. In brief, it appears that when lower primates are in the wild, and fending for themselves, their behaviour is 'civilized', certainly in comparison with their behaviour when they are confined and cared for, which is when they most behave 'like animals'.

Our observations of children lead us to believe that much the same is true of our own species. We have noticed that when children are herded together in the playground, which is where the educationalists and the psychologists and the social scientists gather to observe them, their play is markedly more aggressive than when they are in the street or in the wild places. At school they play 'Ball He', 'Dodge Ball', 'Chain Swing', and 'Bull in the Ring'. They indulge in duels such as 'Slappies', 'Knuckles', and 'Stinging', in which the pleasure, if not the purpose, of the game is to dominate another player and inflict pain. In a playground it is impracticable to play the free-ranging games like 'Hide and Seek' and 'Relievo' and 'Kick the Can', that are, as Stevenson said, the 'well-spring of romance', and are natural to children in the wastelands. Often, when we have asked children what games they played in the playground we have been told 'We just go round aggravating people.' [snip; more descriptions of abusive games]

Such behaviour would not be tolerated amongst the players in the street or the wasteland; and for a long time we had difficulty reconciling these accounts with the thoughtfulness and respect for the juvenile code that we had noticed in the quiet places. Then we recollected how, in our own day, children who had seemed unpleasant at school (whose term-time behaviour at boarding school had indeed been barbarous), turned out to be surprisingly civilized when we met them in the holidays. We remembered hearing how certain inmates of institutions, and even people in concentration camps during the war, far from having a feeling of camaraderie, were liable to seek their pleasure in making life still more intolerable for those who were confined with them [...].

2.2. We own people, and decide who you're with.

You're segregated by age, and divided in classes. Maybe you'd've learned to learn from a kid a couple years older than you, or learned to teach a kid a couple years younger than you. But that's hypothetical, because we've decided you're not to be with those people.

2.3. We own Import.

In standard schooling, kids aren't around adults in adult environments doing adult activities for adult reasons. There's the real world, the adult world, where everything of Import is, and then there's the kid world, which has to make way for the adult world.

From Children's Games in Street and Playground, Opie and Opie, 1969: [IPFS pdf link](#)

What is curious about these embroilments is that children always do seem to have been in trouble about the places where they played. In the nineteenth century there were repeated complaints that the pavements of London were made impassable by children's shuttlecock and tipcat. In Stuart times, Richard Steele reported, the vicinity of the Royal Exchange was infested with uninvited sportsmen, and a beadle was employed to whip away the "unlucky Boys with Toys and Balls". Even in the Middle Ages, when it might be supposed a meadow was within reach of every Jack and Jill in Britain, the young had a way of gravitating to unsuitable places. In 1332 it was found necessary to prohibit boys and others from playing in the precincts of the Palace at Westminster while Parliament was sitting. In 1385 the Bishop of London was forced to declaim against the ball-play about St. Paul's; and in 1447, away in Devonshire, the Bishop of Exeter was complaining of 'yong Peple' playing in the cloister, even during divine service, such games as 'the toppe, queke, penny prykke, and most atte tenys, by the which the walles of the saide Cloistre have be defowled and the glas wyndowes all to brost'.

Should such persistent choice of busy and provocative play-places alert us that all is not as appears in the ghettos of childhood? Children's deepest pleasure, as we shall see, is to be away in the wastelands, yet they do not care to separate themselves altogether from the adult world. In some forms of their play (or in certain moods), they seem deliberately to attract attention to themselves, screaming, scribbling on the pavements, smashing milk bottles, banging on doors, and getting in people's way. A single group of children were able to name twenty games they played which involved running across the road. Are children, in some of their games, expressing something more than high spirits, something of which not even they, perhaps, are aware? No section of the community is more rooted to where it lives than the young. When children engage in 'Last Across' in front of a car is it just devilment that prompts the sport, or may it be some impulse of protest in the tribe? Perhaps those people will appreciate this question most who have asked themselves whether the convenience of motorists thrusting through a town or village is really as important as the well-being of the people whose settlement it is, and who are attempting to live their lives in it.

Let yong Peple go!

From John Taylor Gatto, Dumbing Us Down, 1992: [IPFS pdf link](#)

In Monongahela by that river everyone was my teacher. Daily, it seemed to a boy, one of the mile-long trains would stop in town to take on water and coal, or for some mysterious reason; the brakeman and engineer would step among snot-nosed kids and spin railroad yarns, let us run in and out of boxcars, over and under flatcars, tank cars, coal cars, and numbers of other specialty cars whose function we memorized as easily as we memorized enemy plane silhouettes. Once a year, maybe, we got taken into the caboose that reeked of stale beer to be offered a bologna-on-white-bread sandwich. The anonymous men lectured, advised, and inspired the boys of Monongahela — that was as much their job as driving the trains.

Sometimes a riverboat would stop in mid-channel and discharge a crew who would row to shore, tying their skiff to one of the willows. That was the excuse for every rickety skiff in the twelve-block-long town to fill up with kids, pulling like Vikings, sometimes with sticks instead of oars, to raid the "Belle of Pittsburgh" or "The Original River Queen." Some kind of natural etiquette was at work in Monongahela. The rules didn't need to be written down; if men had time they showed boys how to grow up. We didn't whine when our time was up: men had work to do — we understood that and scampered away, grateful for the flash of our own futures they had had time to reveal, however small it was.

The world isn't yours, it's the adults's. We're driving in our cars on our way to and from important things, and you better get out of the way. They're very important things, they aren't for you, go play (somewhere else).

Imagining that your activities, explorations, and questions in the classroom could be taken up with the solemn seriousness, open reality, and pivotal consequence of a factory floor, a judge's courtroom, or an artist's studio, is very cute and childlikeish. That I would *believe* you, is laughable, though I will of course humor you. If the work I give you is comically fake, well did you expect it to be real? You're only a child.

3. Preference falsification and double binds.

(See [Wiki: Double bind](#), h/t Michael Vassar.)

[Slavoj Žižek:](#)

It's Sunday afternoon. My father wants me to visit our grandmother. Let's say my father is a traditional authority. What would he be doing? He would probably tell me something like, "I don't care how you feel; it's your duty to visit your grandmother. Be polite to her and so on." Nothing bad about this I claim because I can still rebel and so on. It's a clear order. But what would the so-called post-modern non-authoritarian father do?

I know because I experienced it. He would have said something like this, "You know how much your grandmother loves you, but nonetheless I'm not forcing you to visit her. You should only visit her if you freely decide to do it." Now every child knows that beneath the appearance of free choice there is a much stronger pressure in this second message. Because basically your father is not only telling

you, you must visit your grandmother, but you must love to visit it. You know he tells you how you must feel about it. It's a much stronger order.

Your attention is yours (and you must give it to me). Follow your interests (and be interested in what we're "teaching"). You want to be good, right? So you want to follow the rules we make, right? And they're rightful rules, aren't they, or why else would you want to follow them? Develop your own unique specialness (make sure it's one of the things on this list though). God help you if you question our authority to deny you bodily autonomy, and you'd better pretend that CAN and MAY are meaningfully different here.

[Always smile. Refrain from looking out of the window.](#)

Possibilities

Many of these harms can be alleviated or avoided by just not being crazy. (Yes it's that easy.) It's fine if the kids aren't paying attention to what you're teaching, why are you trying to teach 20 kids at once anyway? If something is truly repulsive to a kid's attention, then you're just wrong about what's good for the kid right now, period. And so on.

What's left is for teachers (or "mentors" or "guides" or something) to bring the world to the children. Montessori wrote about this at length. In general, there's much work to be done with teaching; but this is mostly unknown territory, since teachers have mostly so far been obliged to do something other than facilitate learning, and teachers will have to learn how to let the students learn, which is a detailed activity with unknown challenges. Maximize blocks of uninterrupted time. Maximize the environment for opening up the world, deferring to children's interests. Allow the children to make their own environment, like the people that they are.

It would be trivially easy to make school a better product for parents. Have it run until 1720 or later, so parents with work have daycare (kids of course need daycare until they're 18; or at least, that is the revealed preference for whatever reason). Allow flexible sign-in and sign-out. Be completely open to parents visiting. Let kids play outside for many hours, so they aren't stressed, sick, and depressed. Never give homework, so kids can be with their families.

I spoke with someone who runs a pre-school in the spirit of Montessori. Ze told me that ze started a Montessori school, but regulations and money problems made it go under. Ze was sure there's a market, though. I suspect (not having carefully evaluated things) that Effective Altruism is severely underweighting the value of investing in education. I think that harming kids makes them grow up to be more likely to harm other people and be less creative. School almost certainly doesn't matter if AGI comes in the next decade or two, but if we have longer, then more rolls of the dice for brilliant natural philosophers seems like maybe a pretty good way to spend resources.

Contact me at my gmail address, username "tsvibtcontact", if you want to discuss with me possible interventions and funding (not that I'm well-suited to these tasks or have too much energy for them, but I'm interested).

[Not sure Becky's kidding.](#)

Voting Results for the 2020 Review

Full voting results [here](#). Original [2020 Review announcement here](#).

That's it folks! The votes are finalized! The [Annual Review of 2020](#) has come to a close. So ends this yearly tradition that we use to take stock of the progress made on LessWrong, and to provide reward and feedback to the writers and researchers who produced such great works.

Donate to thank the authors (matching funds until Feb 15th 11:59pm)

Speaking of reward and feedback, this year we're doing something new with the Review. Like normal, the LessWrong team will awarding prizes to top posts. But this year we'll be allocating prizes from two different pools of money - the Review Vote pool, and the [Unit of Caring](#) pool.

For each pool, the review panel will be using moderator discretion. We'll be ensuring the prizes go to posts which we believe further our cause of developing the art of rationality and intellectual progress. But for the Review Vote prize pool, our judgment will be strongly informed by the results of the vote. For the Unit of Caring prize pool, our judgment will strongly be informed by the opinions expressed by donors who contribute to the prize pool.

For the Review Vote prize, we will allocate \$10,000.

For the Unit of Caring prize, we will allocate *up to* \$5000, matching the total amount that other LessWrong users contribute to the pool. (i.e. if LessWrong users donate \$4000, the pool will be \$8000. If users donate \$6000, then the total prize pool will be \$11,000).

[Update: the donation period is now over]

If you want to donate while signaling support for particular posts, you can do so using the buttons for individual posts further down the page. Here is your opportunity to not just spend internet points, but to actually spend a costly signal of support for the authors and posts you found valuable!

Donations must be made by February 10th to contribute to the matching pool.

EDIT: deadline extended to the end of February 15th

Complete Voting Results (1000+ Karma)

You can see more detailed results, including non-1000+ karma votes, [here](#).

A total of 400 posts were nominated. 121 got at least one review, bringing them into the final voting phase. 211 users cast a total of 2877 votes. Users were asked to vote on posts they thought made a significant intellectual contribution.

Voting is visualized here with dots of varying sizes (roughly indicating that a user thought a post was "good" "important", or "extremely important"). Green dots indicate positive votes. Red indicate negative votes. You can hover over a dot to see its exact score.



Results

Here are the posts. Note that the donation buttons don't go directly to post authors – they are granted to the Unit of Caring prize pool. The LessWrong moderation team will be exercising some judgment, but the distribution will likely reflect the distribution of donor recommendations.

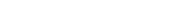
| | | |
|----|---|--|
| 0 | Draft report on AI timelines Ajeya Cotra | |
| 1 | An overview of 11 proposals for building safe advanced AI evhub | |
| 2 | When Money Is Abundant, Knowledge Is The Real Wealth johnswentworth | |
| 3 | microCOVID.org: A tool to estimate COVID risk from common activities catherio | |
| 4 | Alignment By Default johnswentworth | |
| 5 | The Solomonoff Prior is Malign Mark Xu | |
| 6 | Seeing the Smoke Jacob Falkovich | |
| 7 | Pain is not the unit of Effort alkjash | |
| 8 | The ground of optimization alexflint | |
| 9 | Simulacra Levels and their Interactions Zvi | |
| 10 | What Money Cannot Buy johnswentworth | |
| 11 | | |

- [AGI safety from first principles: Introduction](#)
[Richard Ngo](#)
- [The Pointers Problem: Human Values Are A Function Of Humans' Latent Variables](#)
[johnswentworth](#)
- [Coordination as a Scarce Resource](#) [johnswentworth](#)
- [Inaccessible information](#) [paulfchristiano](#)
- [Cortés, Pizarro, and Afonso as Precedents for Takeover](#) [Daniel Kokotajlo](#)
- [My computational framework for the brain](#)
[Steven Byrnes](#)
- [Introduction to Cartesian Frames](#) [Scott Garrabrant](#)
- [Inner Alignment: Explain like I'm 12 Edition](#)
[Rafael Harth](#)
- [Against GDP as a metric for timelines and takeoff speeds](#) [Daniel Kokotajlo](#)
- [The Road to Mazedom](#) [Zvi](#)
- [Anti-Aging: State of the Art](#) [JackH](#)
- [Interfaces as a Scarce Resource](#) [johnswentworth](#)
- [Most Prisoner's Dilemmas are Stag Hunts; Most Stag Hunts are Schelling Problems](#) [abramdemski](#)
- [An Orthodox Case Against Utility Functions](#) [abramdemski](#)
- [Is Success the Enemy of Freedom? \(Full\)](#) [alkjash](#)
- [CFAR Participant Handbook now available to all](#) [Duncan Sabien](#)
- [Introduction To The Infra-Bayesianism Sequence](#) [Diffractor](#)
- [Radical Probabilism](#) [abramdemski](#)
- [Reality-Revealing and Reality-Masking Puzzles](#) [AnnaSalamon](#)
- [Why haven't we celebrated any major](#)

- achievements lately? jasoncrawford
- Some AI research areas and their relevance to
31 existential safety Andrew_Critch
- 32 Search versus design alexflint
- "Can you keep this confidential? How do you
33 know?" Raemon
- 34 Discontinuous progress in history: an update
KatjaGrace
- 35 The Treacherous Path to Rationality Jacob Falkovich
- 36 How uniform is the neocortex? zhukeepa
- 37 To listen well,_get curious benkuhn
- 38 Motive Ambiguity Zvi
- 39 The Felt Sense: What, Why and How Kaj_Sotala
- 40 Choosing the Zero Point orthonormal
- 41 The First Sample Gives the Most Information
Mark Xu
- 42 Nuclear war is unlikely to cause human extinction
landfish
- Why Neural Networks Generalise, and Why They
43 Are (Kind of) Bayesian Joar Skalse
- 44 Can crimes be discussed literally? Benquo
- 45 Credibility of the CDC on SARS-CoV-2 Elizabeth
- Swiss Political System: More than You ever
46 Wanted to Know (I.) Martin Sustrik
- 47 Studies On Slack Scott Alexander
- The Alignment Problem: Machine Learning and
48 Human Values rohinmshah
- 49 Transportation as a Constraint johnswentworth
- 50 The date of AI Takeover is not the day the AI

- takes over [Daniel Kokotajlo](#)
- 51 A non-mystical explanation of "no-self" (three characteristics series) [Kaj_Sotala](#)
- 52 Covid-19: My Current Model [Zvi](#)
- 53 Possible takeaways from the coronavirus pandemic for slow AI takeoff [Vika](#)
- 54 Elephant seal [KatjaGrace](#)
- 55 The Bayesian Tyrant [abramdemski](#)
- 56 Classifying games like the Prisoner's Dilemma [philh](#)
- 57 Have epistemic conditions always been this bad? [Wei_Dai](#)
- 58 Clarifying inner alignment terminology [evhub](#)
- 59 Inner Alignment in Salt-Starved Rats [Steven Byrnes](#)
- 60 Persuasion Tools: AI takeover without AGI or agency? [Daniel Kokotajlo](#)
- 61 Simulacra and Subjectivity [Benquo](#)
- 62 Wireless is a trap [benkuhn](#)
- 63 What happens if you drink acetone? [dynamight](#)
- 64 Subspace optima [Chris van Merwijk](#)
- 65 Create a Full Alternative Stack [Zvi](#)
- 66 Babble challenge: 50 ways of sending something to the moon [jacobjacob](#)
- 67 My slack budget: 3 surprise problems per week [Raemon](#)
- 68 A tale from Communist China [Wei_Dai](#)
- 69 "No evidence" as a Valley of Bad Rationality [adamzerner](#)
- 70 Protecting Large Projects Against Mazedom [Zvi](#)
- 71

- [Extrapolating GPT-N performance](#) [Lanrian](#)
- 72 [Five Ways To Prioritize Better](#) [lynettebye](#)
- 73 [Give it a google](#) [adamzerner](#)
- 74 [Negative Feedback and Simulacra](#) [Elizabeth](#)
- 75 [Coronavirus: Justified Practical Advice Thread](#) [Ben Pace](#)
- 76 [Crisis and opportunity during coronavirus](#) [jacobjacobjacob](#)
- 77 [A Significant Portion of COVID-19 Transmission Is Presymptomatic](#) [jimrandomh](#)
- 78 [How To Fermi Model](#) [habryka](#)
- 79 [Why Artists Study Anatomy](#) [Sisi Cheng](#)
- 80 [Market-shaping approaches to accelerate COVID-19 response: a role for option-based guarantees?](#) [DerekF](#)
- 81 [Authorities and Amateurs](#) [jefftk](#)
- 82 [What counts as defection?](#) [TurnTrout](#)
- 83 [Range and Forecasting Accuracy](#) [niplay](#)
- 84 [Attainable Utility Preservation: Empirical Results](#) [TurnTrout](#)
- 85 [Seemingly Popular Covid-19 Model is Obvious Nonsense](#) [Zvi](#)
- 86 [Training Regime Day 8: Noticing](#) [Mark Xu](#)
- 87 [Rereading Atlas Shrugged](#) [Vaniver](#)
- 88 [Mazes Sequence Roundup: Final Thoughts and Paths Forward](#) [Zvi](#)
- 89 [The 300-year journey to the covid vaccine](#) [jasoncrawford](#)
- 90 [Conflict vs. mistake in non-zero-sum games](#) [Nisan](#)
- 91 [Shuttling between science and invention](#) [jasoncrawford](#)

- 92 [Reveal Culture](#) [MalcolmOcean](#)  
- 93 [Why indoor lighting is hard to get right and how to fix it](#) [Richard Korzekwa](#) 
- 94 [Exercises in Comprehensive Information](#) 
- 94 [Gathering](#) [johnswentworth](#) 
- 95 [The Oil Crisis of 1973](#) [Elizabeth](#) 
- 96 [A Personal \(Interim\) COVID-19 Postmortem](#) [Davidmanheim](#)  
- 97 [How to teach things well](#) [Neel Nanda](#) 
- 98 [What are good rationality exercises?](#) [Ben Pace](#) 
- 99 [Spend twice as much effort every time you attempt to solve a problem](#) [Jsevillamol](#) 
- 100 [How Long Can People Usefully Work?](#) [lynnettebye](#) 
- 101 [Kelly Bet on Everything](#) [Jacob Falkovich](#)  
- 102 [What's the best overview of common Micromorts?](#) [Raemon](#) 
- 103 [GPT-3: a disappointing paper](#) [nostalgebraist](#)  
- 104 [What is meant by Simulcra Levels?](#) [Chris Leong](#) 
- 105 [How to Escape From Immoral Mazes](#) [Zvi](#)  
- 106 [Developmental Stages of GPTs](#) [orthonormal](#) 
- 107 [The case for lifelogging as life extension](#) [Matthew Barnett](#) 
- 108 [100 Tips for a Better Life](#) [Ideopunk](#)  
- 109 [Taking Initial Viral Load Seriously](#) [Zvi](#) 
- 110 [Luna Lovegood and the Chamber of Secrets - Part 1](#) [lsusr](#)  
- 111 [The Reasonable Effectiveness of Mathematics or: AI vs sandwiches](#) [Vanessa Kosoy](#) 
- 112

[Using a memory palace to memorize a textbook.](#)

..

[AllAmericanBreakfast](#)

[The Skewed and the Screwed: When Mating](#)

113



[Meets Politics](#) Jacob Falkovich

114

..

[Zen and Rationality: Just This Is It](#)

[G Gordon Worley III](#)

115

..

[Can we hold intellectuals to similar public](#)

[standards as athletes?](#) ozziegooen

116



[The Four Children of the Seder as the Simulacra](#)

[Levels](#) Zvi

117



[What are some beautiful, rationalist artworks?](#)

[jacobjacob](#)

118



[The Best Virtual Worlds for "Hanging Out"](#) Raemon

119



[Embedded Interactive Predictions on LessWrong](#)

[Amandango](#)

120



[Covid 12/24: We're F***ed, It's Over](#) Zvi

That's all (for now)

Over the next couple weeks the LessWrong team will look over the voting results, and begin thinking about how to aggregate the winning posts into the Best of LessWrong Collection.

Thanks so much to every who participated – the authors who originally wrote excellent posts, the many reviewers who gave them a lot of careful consideration, and the voters who deliberated.

The Long Long Covid Post

A while back I mentioned I'd aim to write a longer post on Long Covid and [Katja Grace's post](#) on it. This is that post. First I deal with Katja's post, then Scott Alexander's [Long Covid: Much More Than You Wanted to Know](#).

My core model of Long Covid after writing this post:

1. Long Covid is real, but less common than many worry it is.
2. Reports of Long Covid are often people who have symptoms, then blame them on Long Covid whether or not they even had Covid. The exception is loss of taste and smell.
3. Long Covid severity and risk is proportional to Covid severity and risk.
4. If you didn't notice you had Covid, you're at very very low risk for developing Long Covid.
5. Vaccination is thus highly but incompletely protective against Long Covid.
6. Children are thus at minimal risk.
7. Omicron is thus less likely to cause serious Long Covid than Delta.
8. My current estimate of the forward-looking-practical-use chance of a healthy non-elderly person getting serious, life-impacting Long Covid from a case of Omicron is about 0.2%, or 1 in 500. This number will decline further once Paxlovid is readily available.
9. Long Covid remains the primary downside of contracting Covid while young and healthy.
10. Diseases often have long-term negative health effects. Long Covid is not fundamentally so different from [Long Other Disease](#). If you are worried going forward about Long Covid you should consider things like permanently not living in a city to avoid diseases.
11. A lot of people are in poor health. It is likely worthwhile to treat your health a lot more seriously than most people do, irrespective of Covid.
12. The Precautionary Principle carries some weight in all this.
13. Remember that the chance of preventing a Covid case via additional Covid prevention, going forward, even with extreme measures, is not all that high.
14. If you compare the potential costs of Long Covid to the costs of Long Covid Prevention, it is obvious the second is a bigger threat.
15. Short-term additional vigilance is reasonable but rapidly becoming less reasonable.
16. Using Long Covid as a reason for not returning to normal once case levels come down would not be reasonable.

Katja's Post

This may or may not be entirely fair, but I am going to use [Katja Grace's post on LessWrong](#) as a steelman of the case for worrying about Long Covid. It is by a thinker I respect, and is clearly advocating for the side of 'be worried,' and seems to aim to be exhaustive. It doesn't cover every concern I've heard, but it's a lot of them.

I'm placing sufficient stock by her selection process that I will focus in detail on the ones she includes.

Katja assigns letters A-R to her points, and I deal with these in order. Before I begin, I'll summarize my takeaways so you can decide which sections you want to read or skip, and have perspective on what you are reading, if you want to go in non-blind.

A, B, C and D establish that Long Covid exists but don't make the case about frequency or causation that they might seem to be making. They do establish that a lot of people have a lot of chronic health conditions.

E is effectively punted to the section on Scott's post since it talks about Scott's estimate.

F, G and H are a claim that there is a bunch of 'dark matter' style damage being done and that it manifests in lots of additional deaths that aren't attributed to Covid. The population statistics don't match up with this, and I consider these to be selection effects or otherwise non-causal.

I reminds us that lesser outcomes can also be concerning but I don't see how they can be that big a share of the overall problem.

J asks about future unknowns. Given how much time has now passed, and what I see as the relevant reference classes, I don't think we need worry about this going forward, but precautionary principle does apply.

K deals with the French study that says that Long Covid is correlated with *thinking* you had Covid but not with Covid itself once you control for whether you think you had it, but when you only check Covid versus symptoms you still find the correlation. I agree with Katja's claim that mostly this isn't psychosomatic, but I do continue to think a lot of it is 'blame whatever is wrong with me on Covid.'

L, M, N and O say already established true things but those facts don't seem to provide evidence for Long Covid being concerning.

P actively seems to go the other way.

Q doesn't seem relevant to the calculations that matter.

R has logic I disagree with, and to extent it was a consideration I think it is no longer one.

The key question with all her points is what you'd expect to find in various different worlds that have different severities of Long Covid.

A. Really bad anecdotes aren't hard to find.

This is strong evidence in the sense that if we *couldn't* easily find really bad anecdotes, we could be confident that Long Covid *wasn't* a Thing worth worrying about.

It's also strong evidence that people *believe* that Long Covid is a thing.

Beyond that, it's *not* strong evidence, because a large percentage of people have had Covid, more people than we realize get long-term debilitating health problems from a range of causes, and anyone who experiences long-term debilitating health problems during the pandemic is likely to consider blaming them on Covid.

Similar things happen for example with Lyme disease. The degree to which Long Lyme is a real thing is similarly disputed, but there's little doubt that a lot of people will claim they're suffering from Long Lyme who very clearly are not.

Another key question is *where the anecdotes are drawn from*. It is not surprising that Katja can receive a response to a bat signal put up on the internet – her social vicinity matters more, so I'm more interested in B than A.

B. Bad anecdotes are common enough to show up in my vicinity

Katja uses the plural in the title. She cites two anecdotes here, [that of Michael Osborne](#) from her extended network, and the claim of a distant relative. As she notes, the distant relative was sufficiently distant that she didn't hear about them having Covid first, and only mentioned it due to the Long Covid discussion, which is a far larger network and involves a much lower reporting threshold.

It seems that Michael ended up recovering, so it wasn't lifelong, but this sounds a lot like losing a year and a half of one's life. It's really bad, bad enough that his account went viral on Twitter.

For reference later, Michael got Long Covid in March 2020, so on average it was more likely severe, wasn't treated as well as we can now, and he was certainly unvaccinated.

So the question becomes, how many people does Katja know as well or better than Michael, and also to what extent knowing Michael caused Katja to write the post and have her concerns.

As an additional data point, in my fully extended networks, I have two examples of Long Covid. One is *also* from someone who was of a broadly similar age, in excellent shape and did a ton of exercise. Not as bad as Michael, but a major hit to their long-term health that they strongly believe is due to Covid. The other is someone older, and I don't know what happened there beyond a vague claim, it's a very weak connection and we haven't communicated in almost two years now.

Also noteworthy is that both of those cases were from March 2020.

If I was quickly ranking how striking hearing about a given person's case of Long Covid would be to me, I'm guessing the younger of these cases was somewhere between 50th and 200th most potentially salient, and the older one between 200th and 1000th. I don't know how many people fall into 'I would have heard about it at all' but I'm guessing the number is in the low thousands.

There's also a third potential case, but the person in question already had other chronic conditions, and claims to have had Covid three times but never got a confirmed Covid diagnosis. I am skeptical.

Other than those cases, I don't know of anyone claiming they have serious Long Covid that I know personally.

So overall, this seems like evidence that things that seem like Long Covid do happen, but it doesn't seem especially strong or scary with respect to frequency.

C. Rates of ambiguously-maybe-quite-bad symptoms seem very high, even for people who only had mild Covid

[This](#) Norwegian study, n = 70k, has, for mild cases (in a sample mixing people who had Covid 1-6 months prior with people who had it 11-12 months prior):

- 10.2% with fatigue (**6.4% higher** than control in the 11-12 months case),
- 7.1% with poor memory (**3.5% higher** than previous control),
- 9.2% with brain fog (**5.3% higher** than previous control).
- 6.9% with shortness of breath (**5.6% higher** than previous control)

Huge if true.

That's among the unvaccinated – she estimates based on other studies that vaccination cuts this in half.

As I said when I first saw this study, these results would show up at the population level. The United States has had 61 million *official* Covid cases, and at least two thirds or so of cases are missed. The *majority* of Americans have had Covid. A lot of those are from December 2021 and January 2022 so it's too soon to know their long-term outcomes, but even excluding them we're looking at half the population, so going from 4% to 7% fatigue, 1.5% to 4% shortness of breath and 4% to 7% brain fog, if it was all happening at once and symptoms are frequently long-term or permanent. I believe these are snapshot numbers. But

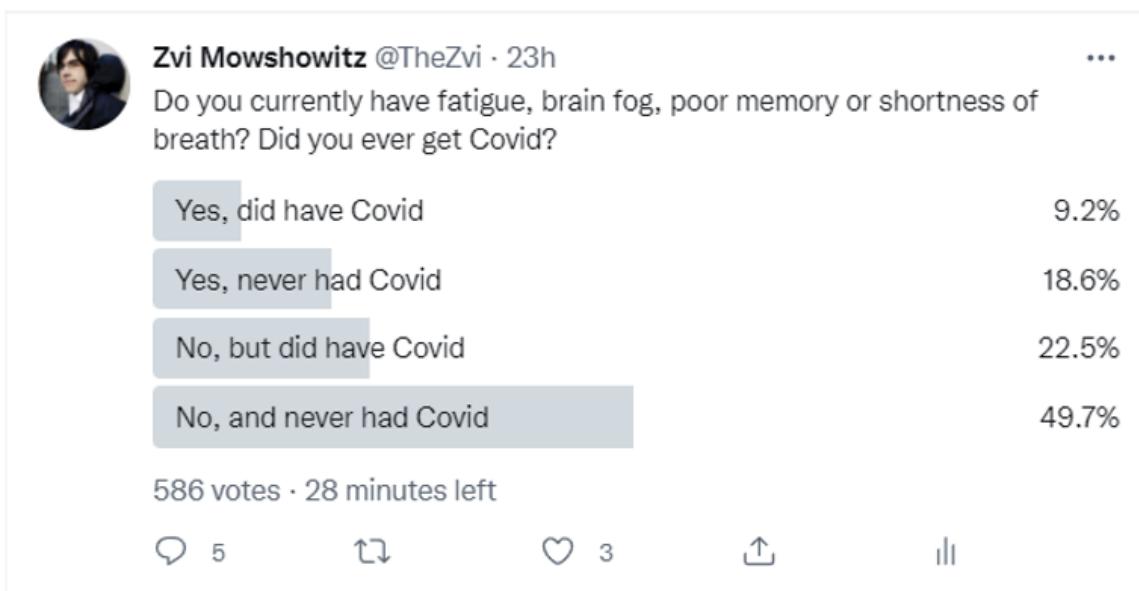
given the correlation between different symptoms is imperfect, and how often the polls I'll get to shortly showed that these symptoms interfere with the ability to work, this is a *very large* set of claims, enough to materially decrease the size of the productive workforce.

Does the anecdata above match several percent increases in a large number of distinct symptoms? No, it doesn't. People are primed by their Covid to look for problems that aren't there, or are barely even there, or were caused by something else.

As Katja notices, these numbers imply that a lot of people can't work.

As an experiment, I did a poll while writing this asking if people (1) had Covid and (2) if they currently have at least one of the four symptoms above – fatigue, poor memory, brain fog or shortness of breath. Every study says there's a correlation between such symptoms and *self-reported* Covid but I'm curious as to magnitude in practice. I expect this to be true regardless of whether there's a substantial causal link, both because the studies say there is one, and because I expect people to be more likely to notice such problems if they have Covid and to think they have had Covid on the basis of now having such problems.

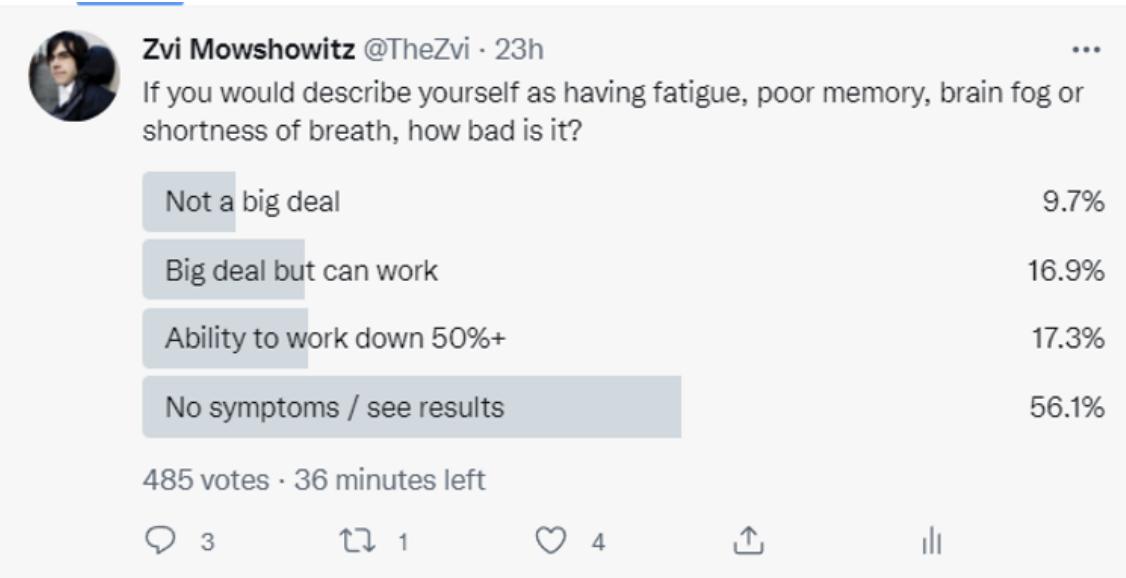
So the real question here is the size of the correlation.



Early results had things reversed a bit, but this makes more sense. Here are the big observations.

1. Wow, that's a *lot* of people reporting those symptoms. Like, a lot. 29%!
2. By comparison, even if there was zero overlap, the Norwegian study said 22%.
3. Of those that report having had Covid, 29% report symptoms.
4. Of those that report not having had Covid, 27% report symptoms.
5. That's 'not significant' in both senses.
6. This is not a scientific study but it also *really is not compatible* with rates doubling or more.

Also, in terms of how big a deal this is:



So what we do know is, these kinds of problems *are a really freaking big deal* in general. Almost 45% of those who report these symptoms say they're wiping out half of their ability to work, and 29% of respondents reported the symptoms in the other poll, so this is at more than 10% of all productivity lost. And for a quarter of all people, at least one of these is being reported as a big deal.

Once again, not a scientific study or representative sample or anything, but ouch. Seems worth exploring this more. Also, I realized I don't know how to thread polls in Twitter's interface, someone in the comments clue me in please for next time.

D. It looks like a lot of people can no longer do their jobs

If a lot of people can indeed no longer do their jobs, and we can be confident this is due to Covid, then this is a big deal.

Katja starts with this, accompanied by bar charts.

Katie Bach of Brookings argued a few days ago that an estimate of 1.1 million people out of work because of long Covid is reasonable, out of 103 million working age Americans she estimated had had Covid, i.e. a roughly 1% chance of being entirely out of work

But that's Katie's estimate and not an actual argument or evidence.

Her second source is [a Washington Post article I'd seen before](#) claiming lots of people were out of work due to this problem.

Hard data is not available and estimates vary widely, but based on published studies and their own experience treating patients, several medical specialists said 750,000 to 1.3 million patients likely remain so sick for extended periods that they can't return to the workforce full time.

Long covid is testing not just the medical system, but also government [safety nets](#) that are not well suited to identifying and supporting people with a newly emerging chronic disease that has no established diagnostic or treatment plan. Insurers are denying coverage for some tests, the public disability system is hesitant to approve many claims, and even people with long-term disability insurance say they are struggling to get benefits.

...

The Washington Post interviewed more than 30 people around the country experiencing the sudden financial slide caused by the long form of the disease.

...

"I have hundreds of patients who have had to take time off for long periods of time, quit their jobs, or get fired from their jobs, or take lesser-paying jobs" because of long covid, said Janna Friedly, vice chair for clinical affairs at the Department of Rehabilitation Medicine at the University of Washington School of Medicine, where she and her team are helping long haulers build strength and return to work.

...

John Buccellato, 64, an emergency medicine doctor at an urgent care clinic on Manhattan's Upper East Side, was hospitalized with the virus in March 2020, in the same hospital where his mother died of covid.

...

Chimere Smith, 39, a middle school teacher in Baltimore who has [testified](#) in Congress about covid's impact on her life, has not worked since she caught the virus in the early weeks of the pandemic.

...

The Social Security Administration said in an email that it has received 16,000 covid-related disability claims since December 2020, but the agency would not disclose how many of those were approved or denied.

...

She said she was laid off in March 2020. She got covid eight months later and has been plagued by fatigue, shortness of breath, joint pain and spikes in body temperature ever since.

...

Hood got sick with covid in October 2020, forcing her this year to close the small business she had run with a friend for 15 years, selling wigs, specialized clothing and other needs for cancer patients. Hood's attempt to return to the job she loved ended in frustration.

...

Two months after the September [2020] birth of her baby, Leon, she tested positive for the coronavirus.

I list all the article's anecdotes both to illustrate that the article is largely a long list of such anecdotes, and also to note the *times of infection*. Two were again very early (March 2020 and 'the early weeks') and the others were all in 2020 as well. To which you could say, sure, it takes time to call Long Covid long and for financial distress to overwhelm you. But there really weren't that many cases in March 2020. What stands out about March 2020 is that cases were on average more severe and we had both poor understanding of how to treat Covid and a lack of hospital capacity. If lots of our examples of Long Covid also come from that early, we should be skeptical that the risks are similar going forward.

Janna Friedly specializes in treating such patients, and they often get referred to her, so her seeing many such patients may not be indicative of a general public health pattern.

Similarly, the ‘many medical specialists’ estimate here is the kind of thing that tends to be inflated for various reasons, and also that doesn’t come with any method, reasoning or data attached.

The only data we get here is the 16,000 disability claims related to Covid since December 2020. Is that a lot? About 500,000 disability claims are filed each quarter, and it’s been four quarters, so that’s about 0.8% of all disability claims. If we say that half the country has had Covid, then if Covid-related disability claims correspond with Covid-related disability, that means getting Covid raises your chances of claiming disability within a year by 1.6%. If some such cases don’t cite Covid, it could be more, and if some people are using Covid as a false justification (either intentionally or unintentionally), it’s less. I can think of other factors as well but the effect here is so small I’m not going to worry about them.

It’s fair to care a *non-zero amount* about something that causes a 1.6% rise in chances of disability within the year, but this does not rise to the level of ‘distort my life for months or years to avoid this.’ If Long Covid was causing several percent of people who get Covid to have a long-term inability to work, also known as disability, it would be a lot more than 1% of new disability claims.

Finally, she offers this:

[This](#) meta-analysis of 81 studies I mentioned earlier also looked at work: “29.0% and 47.4% of those who were employed premorbidly were not able to return to work”; “5% to 90% were unable to reach their pre-COVID employment level” (p. 128) (As noted earlier, a lot of the studies in the meta-analysis seem to be small n, involving hospitalized people, without controls, and I don’t know what they did about this. Also, it’s possible I’m misunderstanding what group the meta-analysis is about, given how crazy high the numbers are).

Feels kind of grim to call these people ‘employed premorbidly.’ Academic writing is so bizarre, and also those are some of the broadest ranges I’ve ever seen and the control problem seems extreme, so it seems like this is a case of having to read the paper.

The results claimed are certainly large and in charge.

The literature search yielded 10,979 studies, and 81 studies were selected for inclusion. The fatigue meta-analysis comprised 68 studies, the cognitive impairment meta-analysis comprised 43 studies, and 48 studies were included in the narrative synthesis. Meta-analysis revealed that the proportion of individuals experiencing fatigue 12 or more weeks following COVID-19 diagnosis was 0.32 (95% CI, 0.27, 0.37; $p < 0.001$; $n = 25,268$; $I^2 = 99.1\%$). The proportion of individuals exhibiting cognitive impairment was 0.22 (95% CI, 0.17, 0.28; $p < 0.001$; $n = 13,232$; $I^2 = 98.0$). Moreover, narrative synthesis revealed elevations in proinflammatory markers and considerable functional impairment in a subset of individuals.

After three months, they’re claiming one third had fatigue and one in five were cognitively impaired. As I’ve noted in the past when discussing many similar studies (some of which likely got into this analysis) proper controls are key. This includes controlling for perception. That all goes double now given my quick survey found such high rates of impairment in the general population at an arbitrary time.

In the studies I’ve previously examined, there were a few that had very strong controls and found little or no Long Covid effect, and a lot that had little or no controls and found strong Long Covid effects.

Looking at their criteria for inclusion, I see the requirement that Covid diagnosis was confirmed, which is good and was not always present in studies I have seen, but I see zero mention of controls of any kind. The list of studies doesn’t mention their controls, nor do the exclusions. So looks like their control strategy was not to have one?

And, well, yeah.

3.3. Methodological quality and risk of bias

Taken together, the NOS rating of the component studies was moderate, evidenced by mean scores of 6.0 out of 9.0 for prospective/ambidirectional cohort studies, 4.1 out of 6.0 for retrospective cohort studies, and 5.6 out of 9.0 for cross-sectional studies.

Common methodological limitations were the failure to include a non-exposed group in cohort studies, failure to ascertain whether outcomes were present prior to COVID-19 infection, and a lack of sample size justification in cross-sectional studies. NOS scores within each category for all component studies organised by design are included (Table S1 in the [supplementary material](#)).

Their discussion section gives one way to interpret all this.

Herein we identified that approximately a third of individuals experienced persistent fatigue and over a fifth of individuals exhibited [cognitive impairment](#) 12 or more weeks following confirmed COVID-19 diagnosis. Similar incidences of fatigue and cognitive impairment, respectively, were observed amongst hospitalized and non-hospitalized populations. Furthermore, in contradistinction to other persistent symptoms which may be self-limiting (e.g., anosmia) ([Hopkins et al., 2020](#)), fatigue and cognitive impairment appear to endure and may potentially worsen over time in susceptible individuals ([Jason et al., 2021](#)), as evidenced by similar proportions of affected individuals at <6 and ≥ 6 months follow-up.

A lower incidence of fatigue and cognitive impairment, respectively, were identified amongst children as compared to adults.

These are two results that don't smell right to me at all.

1. No change in symptom rate over time.
2. No change in symptom rate based on severity or hospitalization.

Those make perfect sense *if the symptoms are unrelated to Covid*. They don't make much sense if this is a Covid-related problem, or match the anecdote where people do mostly get better after a while, which I've heard from a number of sources anecdotally.

They have a good limitations section where they point most of this out as potential issues, including pointing out that changed world conditions could be causing the increase in symptoms, such as via increased depression. The difference is that they have the attitude that it's all fine and don't seem much bothered by not having controls.

Without controls, given all the other data I have, I don't feel like this tells us much we didn't already know.

Some harder to interpret data about long covid sufferers in particular (where I'm not sure how many people count as that) still suggests pretty major issues:

[Matt bell](#) says that [this](#) UK data-set has ~18% of non-hospitalized long covid sufferers with "activities limited a lot."

The way that data set is presented is infuriating – there are tables that list raw counts without reference to the sample size (maybe it's an estimated raw number for the whole country, in which case they're quite small), and tables that are missing the obvious things to be curious about, and so on.

Table 4 estimates that 2.06% of the UK has self-reported Long Covid of any duration on 2 January, and again I presume a majority have had Covid at this point.

Here's an interesting section, the percent is estimated percent reporting Long Covid.

| | | |
|--------------------------|--|------|
| Health/disability status | No health conditions | 1.70 |
| | Activity not limited by health conditions | 2.45 |
| | Activity limited a little by health conditions | 4.24 |
| | Activity limited a lot by health conditions | 5.47 |

I presume this is listing their health conditions before Covid since it makes no sense the other way, but am still somewhat confused.

Here's the start of table 5, including the age distribution since we should put that somewhere.

Table 5. Estimated percentage of people living in private households with self-reported long COVID who first had (or suspected they had) COVID-19 at least 12 weeks previously, UK: four week period ending 2 January 2022

| Domain | Group | Estimate | Lower 95% confidence limit | Upper 95% confidence limit | Percent |
|------------|------------|----------|----------------------------|----------------------------|---------|
| All people | All people | 1.46 | 1.41 | 1.52 | |
| Age group | 2 to 11 | 0.26 | 0.18 | 0.34 | |
| | 12 to 16 | 0.99 | 0.82 | 1.15 | |
| | 17 to 24 | 1.35 | 1.14 | 1.57 | |
| | 25 to 34 | 1.46 | 1.29 | 1.64 | |
| | 35 to 49 | 2.09 | 1.95 | 2.23 | |
| | 50 to 69 | 2.03 | 1.93 | 2.12 | |
| | 70+ | 0.92 | 0.84 | 1.00 | |

The dramatic decline for those over 70 is weird, the death rate isn't *that* high. What's even stranger is this is now people who had Covid over 12 weeks ago, instead of the general population, and the estimate has gone down - 2.06% to 1.46%. And then in Table 6 for 12 months the number is down to 0.86%, seeming to contradict the meta-study that said symptoms don't go away over time, adding to my inclination to dismiss the meta-study as suspected nonsense.

Here is it broken down by symptom, for the general population group with an overall 2.06% estimated base rate of Long Covid.

Table 8. Estimated percentage of people living in private households with self-reported long COVID by symptom and duration, UK: four week period ending 2 January 2022

| Symptom | Any Duration | | | Estin |
|--------------------------------|--------------|----------------------------|----------------------------|-------|
| | Estimate | Lower 95% confidence limit | Upper 95% confidence limit | |
| Weakness/tiredness | 1.04 | 0.99 | 1.08 | |
| Shortness of breath | 0.77 | 0.73 | 0.80 | |
| Loss of smell | 0.76 | 0.72 | 0.79 | |
| Loss of taste | 0.57 | 0.54 | 0.60 | |
| Difficulty concentrating | 0.54 | 0.51 | 0.58 | |
| Muscle ache | 0.52 | 0.49 | 0.55 | |
| Headache | 0.48 | 0.45 | 0.51 | |
| Trouble sleeping | 0.46 | 0.43 | 0.49 | |
| Worry/anxiety | 0.43 | 0.41 | 0.46 | |
| Low mood/not enjoying anything | 0.43 | 0.40 | 0.46 | |
| Cough | 0.43 | 0.40 | 0.46 | |
| Memory loss/confusion | 0.42 | 0.40 | 0.45 | |
| Vertigo/dizziness | 0.31 | 0.29 | 0.33 | |
| Chest pain | 0.29 | 0.26 | 0.31 | |
| Palpitations | 0.25 | 0.23 | 0.27 | |
| Sore throat | 0.22 | 0.20 | 0.24 | |
| Loss of appetite | 0.22 | 0.20 | 0.24 | |
| Nausea/vomiting | 0.16 | 0.14 | 0.18 | |
| Abdominal pain | 0.16 | 0.14 | 0.17 | |
| Diarrhoea | 0.11 | 0.09 | 0.12 | |
| Fever | 0.06 | 0.05 | 0.07 | |

So there's a lot of clustering if this adds up to 2%, and also I'd love to see the control group except there doesn't seem to be one. And always compare to the baseline, for example from Google:

There are **approximately 45 million Americans complaining** of headaches each year. That works out to about one in every six people or 16.54% of the population. More

than eight million Americans visit their doctor for complaints of headache each year.

....

The prevalence of fatigue in the general population has been reported to range from **7% to 45%** (see 1, 2); a recent study found that 38% of US workers reported being fatigued (2).

An additional 1% isn't nothing, but correlation of self-reported claims of that magnitude again does not seem like it should panic us.

So in summary: Do we see evidence of the types of sweeping changes we'd expect to see if several percent of people are suddenly unable to work? No, we don't. We do have newspaper stories about such individuals, but it's plane crash style coverage rather than auto accident style coverage.

I have long been skeptical that such big statistical effects were happening without being noticed more. I continue to be skeptical.

But Katja's next section challenges this.

E. Other people's previous back of the envelope calculations on this are not reassuring.

Matt bell:

"If you're a 35 year old woman, and your risk of ending up with lifelong long COVID from catching COVID is 2.8%, then catching COVID would be the same, statistically speaking, as losing (50 years * 0.18 * 0.028 * 365 days/year) = ~90 days of your life."

Scott Alexander:

"Your chance of really bad debilitating lifelong Long COVID, conditional on getting COVID, is probably somewhere between a few tenths of a percent, and a few percent."

Matt Bell was referencing *the UK data set above* so I have no idea how he can get 2.8%, and in fact my reading of the link says he has it somewhat lower than that but still strangely high.

Scott's post is Scott's, so it deserves more careful attention once I'm finished with Katja's. I agree the conclusions here are not reassuring if we take them at face value, although we've already spent *several percent* of my expected remaining lifespan hiding from this thing, so one could multiply (while remembering that neither is infection fully preventable nor otherwise assured.)

F. Having 'survived' covid looks associated with a 60% increased risk of death (after surviving covid) during the following six months

According to a [massive controlled study published in Nature](#) (more readable summary [here](#)). It also looks like they are saying that this is for non-hospitalized covid patients, though the paper is confusing to me.

Death is different from other things. This is no self-report, also the control in this study is actually trying and the effect size seems large.

The excess death was estimated at 8.39 (7.09–9.58) per 1,000 patients with COVID-19 at 6 months. Individuals with COVID-19 had a higher risk of requiring outpatient care (hazard ratio of 1.20 (1.19–1.21)), at an excess burden of 33.22 (30.89–35.58; all excess burdens are given per 1,000 patients with COVID-19 at 6 months) and at a greater

frequency of 0.47 (0.44–0.49) additional encounters every 30 days (Extended Data Table [2b, c](#)).

As Katja notes, the deaths are noteworthy not only because death (0.8% of the time) but also because dying a lot usually indicates something very wrong otherwise.

And then there's this, which is a question I was going to ask about.

In addition to testing negative-outcome controls (Extended data Table [2a](#)) and to further test the robustness of our approach, we developed and tested a pair of negative-exposure controls. We posited that exposure to influenza vaccination in odd- and even-numbered months between 1 October 2017 and 30 September 2019 should be associated with similar risks of clinical outcomes. We therefore tested associations between exposure to influenza vaccine in even- ($n = 762,039$) versus odd- ($n = 599,981$) numbered months and the full complement of 821 high-dimensional clinical outcomes considered in this study (including all diagnoses, medications and laboratory test results). We used the same data sources, cohort-building algorithm, variable definitions, analytical approach (including weighting method) and outcome specification, as well as a similar length of follow-up and interpretation method. Our results showed that none of the associations met the threshold of significance ($P < 6.57 \times 10^{-5}$) considered in this study (Supplementary Fig. [6](#), Supplementary Tables [22–24](#)).

Incidentally, you have to love when the P value threshold chosen is 0.0000657.

So the same thing *didn't* happen with influenza. That rules out or makes more difficult a bunch of potential explanations that revolve around the controls being inadequate.

The patients had a high rate of [stroke](#) and other [nervous system](#) ailments; [mental health](#) problems such as [depression](#); the onset of [diabetes](#); [heart disease](#) and other coronary problems; [diarrhea](#) and [digestive disorders](#); [kidney disease](#); [blood clots](#); [joint pain](#); [hair loss](#); and general [fatigue](#).

Patients often had clusters of these ailments. And the more severe the case of COVID-19, the higher the chance of long-term health problems, the study said.

So basically *everything*, including hair loss. One naturally suspects that this has to do with Covid having an easier time infecting those in generally poor health in various ways, or Covid having an easier time being sufficiently symptomatic to get noticed.

Note that severity *did* matter here. The more severe the case, the more problems (which would also be true if these were underlying problems being selected for). And since this excludes hospitalizations, that implies the missing cases would bring these averages *up*.

And since we're talking about so many Covid cases, this would need to very much show up on the excess death tables, within an order of magnitude of the Covid deaths themselves.

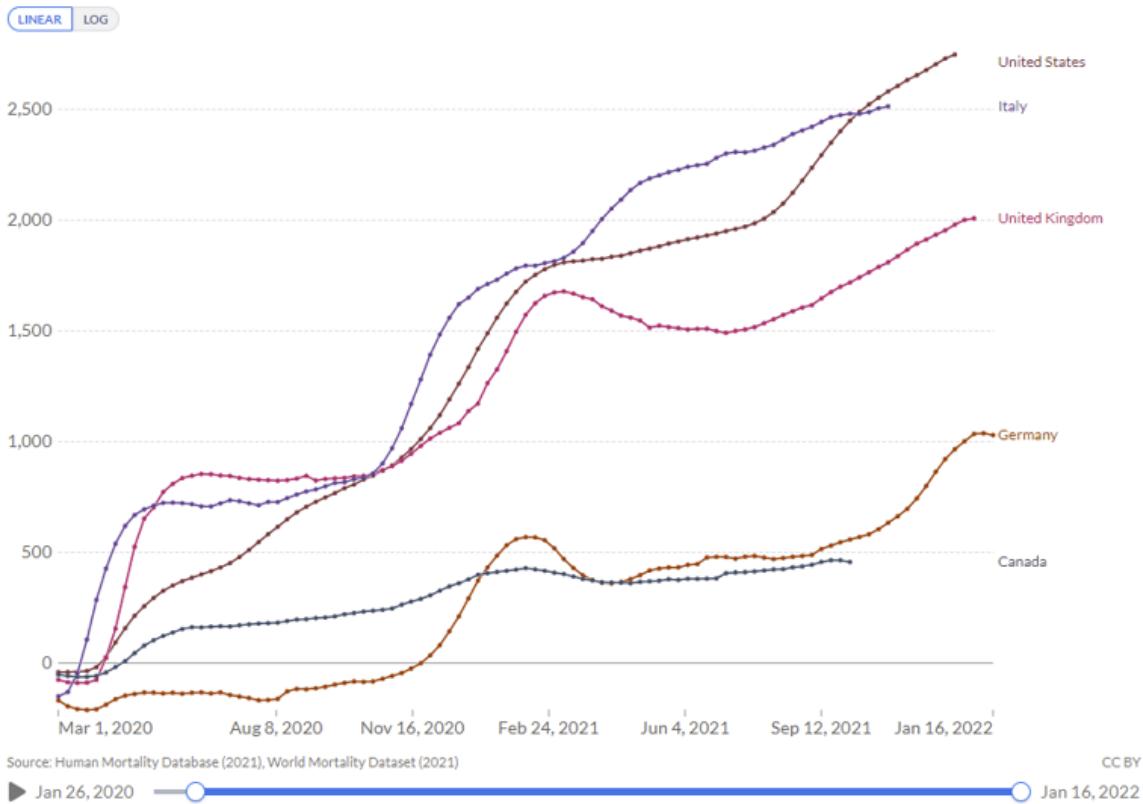
The weird thing is that the magnitude of these effects doesn't seem compatible with the UK estimates or others that say a few percent chance of problems. This is an 0.8% chance of *outright death*, and a lot of other stuff short of that, and the period after six months presumably adds more than that. So if the majority of the nation has had Covid, that would imply over a million excess deaths *on top of recorded Covid deaths*, although perhaps moderately less if we factor in severity.

How much excess mortality did we in fact observe?

Excess mortality: Cumulative number of deaths from all causes compared to projection based on previous years, per million people

Our World
in Data

The cumulative difference between the reported number of deaths since 1 January 2020 and the projected number of deaths for the same period based on previous years. The reported number might not count all deaths that occurred due to incomplete coverage and delays in reporting.



The USA has about 2600 confirmed deaths per million people, and 2740 excess deaths per million people. That does not leave room for this kind of extra death toll, and is exactly what one would expect if this was a null effect, with a small number of missed Covid deaths filling in the gap. This doesn't *rule out* that there's a large effect here, but those extra deaths would have to be compensated for by reduced deaths from other causes missed by the expected death calculation.

One idea I had was to check the *early* figures. It would take time for Long Covid to result in deaths. So we should see more unexplained excess mortality later in the pandemic, and unexplained *missing* mortality early to make the math work. Do we see that?

May 3, 2020: 256 excess deaths per million, 213 Covid deaths per million. The opposite effect, likely due to more unidentified cases very early.

August 2, 2020: 582 excess deaths per million, 466.8 Covid deaths per million. Still the opposite effect.

May 30, 2021: 1903 excess deaths per million, 1783 Covid deaths per million.

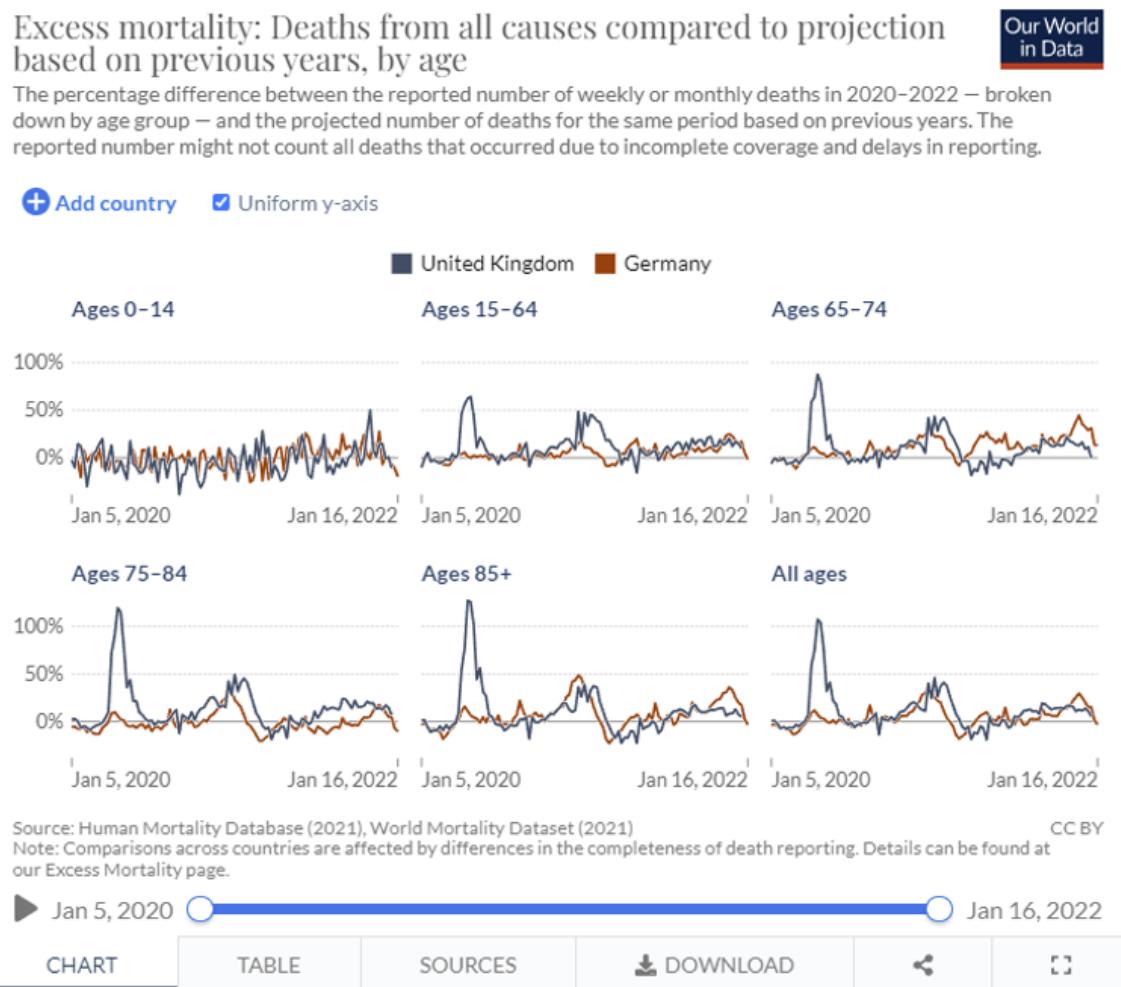
So basically early on there were 'missing' deaths unaccounted for, then deaths kept pace. [The curves look the same](#) with no evidence of lag, or any longer term effects from various earlier waves.

Thus, I defy the data here. There is not a 60% increased risk of death *due to* Covid in the six months after infection. We would see it in the population data. Somehow, there's a statistical

artifact here, this is correlation rather than causation, and the controls don't work right to correct for the problem.

G. Overall deaths from everything have been very unusually high at points in 2021, even in 15-64 age group

[From Our World In Data:](#)



So first off, if you add the United States the 0-14 graph is mostly *below the 0% mark* so let me reiterate *one more time* that kids really, really don't die of Covid.

What about all that excess mortality we see here? What we see are the waves of Covid infections. What we *do not* see are *delayed effects* from those waves, as I noted earlier.

(The Center Square) – The head of Indianapolis-based insurance company OneAmerica said the death rate is up a stunning 40% from pre-pandemic levels among working-age people.

It's because the scale is *the baseline death rate*. Covid kills you at a rate *approximately proportional* to your age-based risk of other death. Early on before vaccines arrived and for that and other reasons death rates fell, a rule of thumb was that it doubled your chance of dying this year. That's why most of the graphs look the same.

A 40% jump in working-age deaths implies a lot of people were getting Covid, but doesn't require there to be lots of mysterious deaths. If we checked only working-age people's excess mortality versus Covid deaths, I bet we don't see any divergence from the patterns above. Note that the quote *doesn't say* how many such folks officially died from Covid, which seems like a weird omission if the quote was supposed to be about Long Covid.

Note especially the final part of the graphs. We do not see excess deaths at all, but a huge portion of people got Covid in December and early January, and many of them never knew it. No visible effect.

This will get much stronger as we get deep into February and then March. If Omicron raises your chances of death afterwards by a lot, *we will know it*.

There's nothing to explain here. This is evidence *against* Long Covid killing people.

H. Sounds like these things involve various damage throughout body, and my guess is that that ends up being regrettable in ways not captured in 'hours lost to fatigue this year'

See [Nature study](#) in F. I also feel like I've seen this a lot, but don't have that many examples immediately on hand. Here's [one other example](#), not ideal because note that these are hospitalized younger people:

For people younger than 65 who were hospitalized with COVID-19, the risk of death in the 12 months after the infection was 233% higher than it was for people who did not have the disease, results published in the journal [Frontiers in Medicine](#) have shown.

Nearly 80% of all deaths of people in the study who had recovered from COVID-19 in the past 12 months were not due to cardiovascular or respiratory causes, suggesting that the impact of the virus is significant and wide-ranging, even after the initial infection has been fought off.

My guess is that all the symptoms are a spectrum, and if the worst looks like an unbelievable amount of cognitive impairment and a pot pourri of organ dysfunctions, or death post-infection, then probably everyone gets a handful of cognitive impairment and organ dysfunction.

There's a wide variety of symptoms being checked for here, not merely fatigue, in a very grab-bag kind of fashion that should pick up on most things.

We also deal with mortality directly above, and the 233% higher number referenced here seems *obviously* driven by selection effects. If you're hospitalized for Covid while non-elderly, that's a huge sign you were already unhealthy, and the effect size here is also way too big for anything else. To say essentially 'oh 80% of deaths after Covid were not from anything Covid seems like it causes, in a similar distribution of causes to deaths of other people, and that simply shows how *nefarious and mysterious* this problem is' ignores the obvious hypothesis, which is that those 80% of deaths have very little to do with Covid. Seems like a case of trapped priors.

I. It's easy to just think about these worst case outcomes, but there are a lot of more probable non-worst case outcomes that would still be a non-negligible bummer.

I see people mostly estimating the worst cases, but my guess is that the more probable non-worst case outcomes (e.g. lesser fatigues and brain fogs, kidney damage, arrhythmias etc), are not nothing.

Sure, agreed as far as it goes. Often, as with Covid prevention, [the dust specks are more worrisome than the torture](#). Many people complain *very loudly* about how bad various dust

specks can be, and I agree that the specks of Covid prevention are the dominant form of damage at this point. However for Long Covid in particular the more severe long hauls seem to be relatively common enough they clearly dominate, and the minor stuff is minor. Not everything is counterintuitive.

J. Future unknowns

Across all diseases, how much of their disutility is obvious in the first two years? Saliently right now: we've had Epstein-Barr for ages and only now noticed that it apparently has a ~1/300 chance of causing MS, usually showing up between 20 and 40, long after the virus, and wreaking brutal destruction. I'm not sure whether we would realize how bad HIV was if it had appeared two years ago and lots of people had it, but nobody had had it for more than two years yet.

I think mostly this isn't right, and also we've previously dealt with a lot of coronaviruses. Yes, HIV wouldn't have shown us how bad it is, but a big hint is that *you don't actually get rid of HIV on your own*. So it's an ongoing infection, which raises our prior that there will be a long-term problem. Epstein-Barr seems like a better parallel, and a ~1/300 chance of that kind of thing is not great, but as far as we know this is pretty unique and we need to divide that by the size of the reference class, at which point it doesn't seem like that big a concern. Also, given Epstein-Barr has been around for ages, we'd have to apply that same logic to every other disease out there.

That doesn't mean I think saying 'Precautionary Principle' is invalid here. When in doubt, 'don't catch the new virus that's going around even if I don't have a specific reason why not' seems like a good principle. But it's infected a *lot* of people over two years, so I think it applies a lot less now than it did earlier, and despite that I still have this carrying a lot of the remaining weight in favor of prevention (beyond vaccinations) for the young and healthy.

K. Long covid probably isn't psychosomatic

A [French study](#) found that long covid is barely associated with having had covid according to an antibody test, yet associated with believing one has had covid (which itself is unrelated to the antibody test results).

At first I (and I think others) thought that if this wasn't some error, then long covid was likely psychosomatic and not caused by physically having covid. But on further thought, that's totally wrong: this pattern could be caused by beliefs causing illness, but it could also be caused by illness causing beliefs, which obviously happens all the time. That is, people's guesses about whether they had covid are heavily influenced by their symptoms.

That second interpretation was my read on the paper as well, although with a non-zero amount of the first one. If the result is real, my guess is it's mostly about people assuming that if they feel fatigue they must have had Covid, rather than primarily people who believe they had Covid therefore thinking themselves into being fatigued.

Neither of those is a reason to worry.

It seems to me that we have other data that basically rules out the possibility that long covid is imaginary (e.g. see [Nature study above](#) on laboratory abnormalities and raised death rate). Though psychosomatic illness is weird - my understanding is that it could in principle still be psychosomatic, while yielding measurable physical laboratory abnormalities, though intuitively I'd be fairly surprised to learn that the same new psychosomatic syndrome had gripped millions in the wake of a particular belief they had, and raised their risk of death by half. Maybe I'm missing something here.

If Long Covid was psychosomatic, that might be an argument *against testing*, but the symptoms are still real. It still counts. To me, the French study is very strong evidence that

Long Covid isn't the big deal it looks like elsewhere, and I already had a very different view than Katja of the other studies – I don't think that risk of death is real, which is both part of why I think Long Covid is rare, and also puts psychosomatic explanations more into play to the extent it matters. But again, I don't think that's a major factor here, and also even if it was, I still say it counts.

In the comments to Katja's post are some objections to the French study, one of which [goes so far as to call it debunked](#) (or in this case bunk, which is oddly the same thing as debunked and the opposite of bunked) because seropositivity is also, in this model, bunk:

That French study is bunk.

Seropositivity is NOT AT ALL a good indicator for having had covid:
https://wwwnc.cdc.gov/eid/article/27/9/21-1042_article

It is entirely possible that all those patients who believe they had COVID are right.

Some researchers believe absence of antibodies after infection is positively correlated with long covid (I don't have a source).

This study is bunk and it's harmful for adequate treatment of seronegative patients. The psychosomatic narrative has been a lazy answer stifling solid scientific research into illnesses that are not well understood yet.

This is not an objection to only the French study, but to the use of seropositivity at all. I don't know of another lab option post-hoc, so it's basically saying all we can do is use the PCR test at the time.

So first off, no, it is *very much not possible* that all those patients are correct, that's obvious nonsense even before I read the paper. If you ask for self-reports then *of course* some number of people are going to report incorrect perceptions, regardless of the accuracy of the test.

It's still important to know if the test is garbage, but before looking at the paper, I'll note that the test result *did correlate* in the study with loss of taste and smell, which means it definitely also correlated with Covid. It did not correlate with other symptoms. So in order for this to be true, *other* long Covid symptoms would need to strongly correlate with the test not working – the test simply not being that accurate is insufficient here even if true.

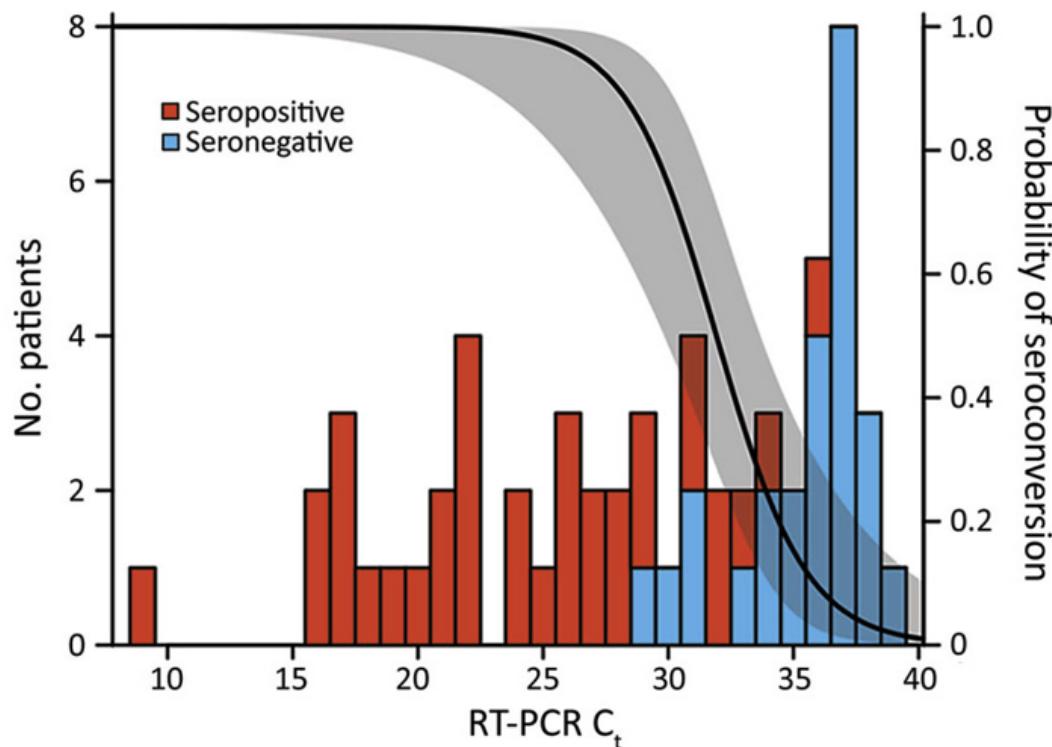
The study found that only 46 of 72 people (not that many but we use what we've got) with positive PCR tests were positive for antibodies three weeks or more afterwards. Being younger or having higher Ct levels made antibodies less likely to be present.

However, seronegative persons were on average 10 (95% CI 3–17) years younger than seropositive persons ([Figure 1](#), panel A) and exhibited RT-PCR Ct values that were 11 (95% CI 8–14) cycles higher ([Figure 1](#), panel B). Moreover, logistic regression showed a precipitous decline in the probability of seroconversion at higher Ct values ([Figure 2](#)). For example, a Ct of 35 predicted only a 15% (95% CI 5%–37%) probability of seroconversion, which decreased further with increasing Ct values. Thus, low nasopharyngeal viral loads seem insufficient to elicit a systemic antibody response.

High Ct levels not leading to persistent antibodies makes sense, as this should be a combination of false positives and mild cases, neither of which seem likely to motivate the body to keep antibodies around. The false positives of course are then 'true negatives' in the antibody test.

The whole effect here is from mild cases. The correlation is no joke, it's more like *the entire ballgame*. Sufficiently mild cases don't show up, non-mild cases always do:

Figure 2



[This FAQ](#) about Ct scores explains that numbers in the 30s can easily not represent a meaningful infection.

7. What can CT values tell us? Samples with CT values <32 generally contain sufficient genetic material for WGS and are more likely to contain replication competent virus. Although there are limitations in the use of CT values, they are one factor to consider when evaluating molecular test results and can be useful in assessing the trend in the viral load. If there is high suspicion of a new infection, laboratories may attempt WGS on samples with CT value <32.

[Or this:](#)

Several studies have correlated Ct values with the presence or absence of infectious virus detectable in culture [3]. One report from Canada's National Microbiology Laboratory (CNML) observed that PCR positive respiratory specimens with Ct values >24 were viral culture negative [4]. In comparison, the United States Centers for Disease Control and Prevention (CDC) reported that attempts to recover SARS-CoV-2 in culture of upper airway samples was generally unsuccessful when their assay Ct values were >35 (unpublished).

Thus, I find the most logical explanation for this 'PCR tests that didn't lead to antibodies were mostly because the person never had Covid and we need to be using a PCR threshold of more like 28-30 and using 40 is just silly.'

Similarly, the French study itself says this about accuracy of the test, [with this source](#) that has much bigger samples:

A test was considered positive for SARS-CoV-2 when the results indicated an optical density ratio of 1.1 or greater (sensitivity, 87%; specificity, 97.5%).

We also have [this analysis](#) from Bucky, which had a critique I find more interesting.

My favorite part isn't the critique though, it's that they gave the serology tests back to the subjects *before asking whether they had Covid* and everyone basically ignored them and said (best imagined with a heavy French accent) 'yes, we saw your test and we choose to ignore it.'

Of everyone who had a positive serology results, only 41.5% replied that they thought they'd had COVID. Of everyone who thought they'd had COVID, 50.4% had had a negative serology result.

So we know from the other study's chart that these are some very stubborn people. If you get a positive serology test in this context, *oh yes you had Covid*. The people with Ct scores that were very high *still* didn't have antibodies (12 of 12 and 17 of 18). The idea that *the majority* of these people still didn't think they had Covid is *super* strong evidence that half the cases were even back then *missed entirely*, and also that such people simply didn't care about the serology results. If you'd shown me a positive serology test in January 2021 and asked me if I'd had Covid, I'd have said yes, because *you just showed me the test, so I guess I must have*.

The French public did not think this way.

I find that to be an interesting study result in its own right. People *simply don't believe* such tests, or don't even care enough to look at them in context.

Check this header out on its own:

Table 2. Descriptive Statistics of Symptom Prevalence by Belief and Serology Test Result Status

| Symptom | Total No. | No. (%) of participants | | Serology+ ^a | | P value ^b |
|---------|-----------|-------------------------|----------------------|------------------------|-------------------|----------------------|
| | | Serology- ^a | Belief- (n = 25 271) | Belief+ (n = 461) | Belief- (n = 638) | |
| | | | | | | |

So people were 1.8% likely to report having been positive if the test was negative versus 41% if the test was positive. 1.8% is lizardman territory.

But Bucky's point is that:

1. Long Covid correlated with reported Covid-19.
2. If you control for *reported* Covid-19, the serology result no longer predicts long Covid.
3. However, if you ignore people's self-reports and simply look at serology versus symptom reports, we still get increased levels of symptoms.
4. It doesn't matter that people have weird ideas about whether they got Covid that are caused by later problems, *you can ignore that*.

Interesting. So model 1 here looks only at beliefs, model 2 at serology, model 3 at both.

Table 3. Associations Between Persistent Symptoms, Belief, and Serology Test Results

Table 3. Associations Between Persistent Symptoms, Belief, and Serology Test Results

| Symptom | No. | Odds ratio (95% CI) ^a | | Model 3 | |
|---------------------------------------|------|----------------------------------|---------------------|---------------------|------------------|
| | | Model 1 Belief | Model 2 Serology | Belief | Serology |
| Sleep problems | 2729 | 1.09 (0.88-1.36) | 0.96 (0.77-1.19) | 1.14 (0.89-1.46) | 0.91 (0.71-1.15) |
| Joint pain | 1894 | 1.32 (1.01-1.71) | 1.03 (0.79-1.35) | 1.39 (1.03-1.86) | 0.89 (0.65-1.21) |
| Back pain | 1630 | 1.41 (1.10-1.80) | 1.16 (0.91-1.49) | 1.40 (1.05-1.85) | 1.01 (0.76-1.33) |
| Digestive tract problems ^b | 909 | 1.92 (1.43-2.57) | 1.06 (0.73-1.50) | 2.19 (1.57-3.06) | 0.73 (0.49-1.08) |
| Muscular pain, sore muscles | 867 | 1.79 (1.29-2.48) | 1.33 (0.94-1.87) | 1.78 (1.22-2.59) | 1.01 (0.68-1.50) |
| Fatigue | 766 | 5.20 (4.20-6.43) | 2.59 (2.03-3.30) | 4.90 (3.79-6.33) | 1.13 (0.84-1.52) |
| Poor attention or concentration | 644 | 3.63 (2.79-4.71) | 2.10 (1.57-2.82) | 3.42 (2.50-4.67) | 1.13 (0.79-1.61) |
| Skin problems | 632 | 1.36 (0.92-2.00) | 0.65 (0.39-1.06) | 1.79 (1.17-2.73) | 0.49 (0.29-0.85) |
| Other symptoms ^c | 514 | 3.07 (2.22-4.25) | 1.91 (1.32-2.75) | 2.93 (1.99-4.31) | 1.10 (0.71-1.70) |
| Sensory symptoms | 492 | 1.60 (1.02-2.51) | 0.77 (0.43-1.38) | 2.06 (1.25-3.40) | 0.54 (0.28-1.03) |
| Hearing impairment | 479 | 1.47 (0.90-2.41) | 1.22 (0.73-2.03) | 1.45 (0.82-2.55) | 1.03 (0.57-1.84) |
| Headache | 360 | 2.52 (1.71-3.73) | 1.69 (1.10-2.59) | 2.40 (1.52-3.80) | 1.10 (0.67-1.82) |
| Breathing difficulties | 256 | 8.16 (5.95-11.19) | 3.60 (2.48-5.24) | 7.75 (5.25-11.43) | 1.11 (0.70-1.76) |
| Palpitations | 213 | 5.27 (3.55-7.82) | 2.61 (1.62-4.19) | 5.14 (3.18-8.29) | 1.05 (0.59-1.87) |
| Dizziness | 178 | 3.23 (1.88-5.56) | 2.37 (1.33-4.24) | 2.71 (1.40-5.24) | 1.42 (0.70-2.88) |
| Chest pain | 174 | 7.34 (4.95-10.88) | 3.70 (2.33-5.87) | 6.58 (4.02-10.75) | 1.25 (0.70-2.22) |
| Cough | 167 | 4.67 (3.00-7.25) | 2.22 (1.25-3.97) | 4.85 (2.75-8.56) | 0.91 (0.45-1.83) |
| Anosmia | 146 | 28.66 (20.16-40.74) | 15.69 (10.85-22.70) | 16.37 (10.21-26.24) | 2.72 (1.66-4.46) |

^a Model 1 includes belief only, controlling for age, sex, educational level, and income. Model 2 includes serology test results only, controlling for age, sex, educational level, and income. Model 3 includes both belief and serology test results, controlling for age, sex, educational level, and income. We additionally tested interactions between serology test results and belief among all of the symptoms, and none were significant.

^b Digestive tract problems refer to the presence of 1 or more of the following

persistent symptoms: nausea, diarrhea, constipation, and stomach pain.

^c Other symptoms refer to additional symptoms that patients declared and are not on the symptom list, plus symptoms with a low number of cases (<100), such as speech problems (n = 56), fever or fever sensation (n = 26), anomaly of the facial nerves (n = 16), and discomfort (n = 12).

In model 2, which ignores belief entirely, we still do get a bunch of increased risks purely from a positive serology result. In every case, belief is a *much better* predictor of reported symptoms, again strongly suggesting the model that symptoms cause belief in having had Covid, and perhaps that belief in having had Covid causes symptoms as well. And once we control for belief, serology tells us *exactly* nothing except for anosmia which is so tightly bound to Covid that the correlation remains even though belief correlates that much better.

Also, there's Table 2 to look at the raw data.

Table 2. Descriptive Statistics of Symptom Prevalence by Belief and Serology Test Result Status

Table 2. Descriptive Statistics of Symptom Prevalence by Belief and Serology Test Result Status

| Symptom | Total No. | No. (%) of participants | | | | P value ^b |
|---------------------------------------|-----------|-------------------------|----------------------|------------------------|-------------------|----------------------|
| | | Serology- ^a | Belief- (n = 25 271) | Serology+ ^a | Belief+ (n = 461) | |
| Sleep problems | 2729 | 2580 (10.4) | 49 (10.9) | 55 (8.7) | 45 (10.1) | .58 |
| Joint pain | 1894 | 1802 (7.3) | 30 (6.7) | 26 (4.2) | 36 (8.2) | .02 |
| Back pain | 1630 | 1525 (6.2) | 32 (7.1) | 33 (5.2) | 40 (9.1) | .048 |
| Digestive tract problems ^c | 909 | 838 (3.5) | 33 (7.4) | 20 (3.3) | 18 (4.2) | <.001 |
| Muscular pain, sore muscles | 867 | 808 (3.2) | 22 (4.8) | 18 (2.9) | 19 (4.3) | .16 |
| Fatigue | 766 | 625 (2.5) | 57 (12.6) | 22 (3.5) | 62 (13.8) | <.001 |
| Poor attention or concentration | 644 | 555 (2.2) | 34 (7.5) | 17 (2.7) | 38 (8.5) | <.001 |
| Skin problems | 632 | 598 (2.4) | 17 (3.8) | 6 (1.0) | 11 (2.5) | .02 |
| Other symptoms ^d | 514 | 463 (2.0) | 17 (3.8) | 8 (1.3) | 26 (6.0) | <.001 |
| Sensory symptoms | 492 | 463 (1.8) | 16 (3.5) | 8 (1.3) | 5 (1.1) | .02 |
| Hearing impairment | 479 | 456 (1.8) | 7 (1.5) | 6 (1.0) | 10 (2.2) | .33 |
| Headache | 360 | 323 (1.3) | 13 (2.8) | 8 (1.3) | 16 (3.6) | <.001 |
| Breathing difficulties | 256 | 192 (0.8) | 29 (6.4) | 9 (1.4) | 26 (5.8) | <.001 |
| Palpitations | 213 | 175 (0.7) | 17 (3.7) | 6 (1.0) | 15 (3.4) | <.001 |
| Dizziness | 178 | 158 (0.6) | 7 (1.5) | 5 (0.8) | 8 (1.8) | .002 |
| Chest pain | 174 | 138 (0.6) | 14 (3.1) | 2 (0.3) | 20 (4.5) | <.001 |
| Cough | 167 | 144 (0.6) | 10 (2.2) | 2 (0.3) | 11 (2.5) | <.001 |
| Anosmia | 146 | 75 (0.3) | 20 (4.4) | 7 (1.1) | 44 (9.9) | <.001 |

^a Serology test result negative (-) or positive (+) for SARS-CoV-2 infection.
^b Reflects the statistical significance of between-group differences according to χ^2 tests.
^c Digestive tract problems refer to the presence of 1 or more of the following persistent symptoms: nausea, diarrhea, constipation, and stomach pain.
^d Other symptoms refer to additional symptoms that patients declared and are not on the symptoms list, plus symptoms with a low number of cases (<100), such as speech problems (n = 56), fever or fever sensation (n = 26), anomaly of the facial nerves (n = 16), and discomfort (n = 12).

PDF

We see some weird stuff here. Conditional on disbelief in having had Covid, positive serology reduces how often people experienced all the common problems except fatigue and poor attention or concentration, with also noticeable increase in breathing difficulties, but that's it. This makes some sense, in that if you both have the issue and also had Covid, then you're really likely to notice you had it, whereas if you didn't actually have it and also have a negative serology test in hand, there are presumably some barriers to fooling yourself.

One thing to note is that there are a number of symptoms where the worst-off group are serology negative but belief positive: Digestive problems, muscular pain, skin problems, sensory symptoms, breathing difficulties. This to me is very strong evidence that the belief-positive group is largely telling a story based on their symptoms.

Model 2 results show that the likelihoods of experiencing the following persistent symptoms are increased by having had COVID (odds ratio / percentage point increase vs serology negative):

- Fatigue (2.59 / 5.0%)
- Anosmia (15.69 / 4.3%)
- Poor attention/concentration (2.10 / 2.8%)
- Breathing difficulties (3.60 / 2.3%)
- Chest pain (3.70 / 1.4%)
- Palpitations (2.61 / 1.2%)
- Headache (1.69 / 0.9%)
- Dizziness (2.37 / 0.6%)
- Cough (2.22 / 0.6%)
- Other symptoms (1.91 / 1.3%)

If we add all the percentage point increases (i.e. how many more percentage points serology positive participants experienced persistent symptoms vs serology negative

participants – data from [table 2](#)) then we get 20.3%. So having COVID on average gives you ~0.2 persistent symptoms vs not having COVID, with presumably some people having more than one symptom.

This is roughly in line with Scott's conclusions in [Long COVID: Much more than you wanted to know](#). The specific symptoms experienced are also in line with that post, so if that post reflects your current understanding of Long COVID then I wouldn't update much based on this study except to add some more confidence to a couple of the points Scott makes:

2. The prevalence of Long COVID after a mild non-hospital-level case is probably somewhere around 20%, but some of this is pretty mild.

3. The most common symptoms are breathing problems, issues with taste/smell, and fatigue + other cognitive problems.

So I will notice that these two 20% chances are not the same. Even if the symptoms were uncorrelated entirely you would expect only 16% of people to have at least one symptom in the French study, and I have zero doubt all of this highly correlates.

That's before you adjust for the percentage of the observed effect that is acting through *belief in having had Covid*, which I don't doubt is doing real work here. And of course, this is all early-pandemic stuff, and of course there's still the thing where people who got Covid during these time frames were substantially different from the people who didn't.

I do find this all convincing to the extent that it shows the French study isn't contradicting the other studies. We should assume that a large percentage of long Covid observations are based on people thinking they had Covid because they have health problems, and all the rest, but the whole thing doesn't simply go away this easily.

L. The general discussion of what is going on with people's brains sounds terrible

I mean, yes, that is a decent summary of half of LessWrong starting from its founding and also the only reasonable conclusion to draw from Twitter or Facebook.

I think I'm about half kidding, people's brains do not work the way we'd like them to all that often.

The [list of different plausible routes to brain damage](#) occurring [from Covid] according to Nature—some brain cells getting broken, some blood restrictions causing little strokes, some immune system attacking your own brain type issues—is one I want very little to do with.

Famously, the worst thing someone with trouble sleeping due to a medical problem can do is type their symptoms into WebMD. Never do this.

Similarly, if someone starts brainstorming around to list all the ways Covid *could* damage your brain directly and indirectly, and describing them in detail, you are *not* going to like it, but this should come as no surprise.

Yes, if these things are real, they are very much not fun. but we knew that already. Things that mess people up mostly sound terrible. If someone described all the things that went wrong with you via *aging*, it would be so much scarier, so once again fund and otherwise support longevity research and all that.

Also, the descriptions here are in hospitalized patients, which is very much not a representative sample of the young. It's scaremongering whether or not that was the intention. Comparing snapshot levels of 'toxic chemicals' in some Covid patients' brains to

levels under Alzheimer's is the kind of cherry-picked, emotionally-loaded comparison you can't let bother you.

M. It sounds like covid maybe persists in your body for ages?

Seems like the virus lives throughout your organs long after recovery, based on autopsies, including of mild/asymptomatic covid sufferers ([summary, paper](#)).

Based on autopsies of people *who died while still infected*, they concluded that it '*can persist for months*' in various places in the body. How could this *possibly* inform us on whether the virus persists after recovery, when no one studied recovered from Covid? This is the kind of 'did you know X can in theory do Y?' that is primarily useful in ghost stories, and seems misleading slash unjustified enough to be on the borders of what is allowed under [Bounded Distrust](#). But something always *can* do something, which Katja then reads as 'maybe persists' in general.

As far as I can tell, no one thinks Covid can spontaneously reemerge after you recover. If no non-humans had Covid, and we cured every human two weeks ago, that would be the end. Again this is not like HIV, where the virus 'persists in the body' in the sense of you never curing the disease.

N. Later rounds of covid are probably bad too

This assumes that later covids are basically free, once you've done it once, in a way that isn't true for e.g. crashing your car. My guess is that later bouts are less bad on average, but far from free.

In [my survey](#), of three people with lasting problems who got covid at least twice, one got the problems with the first, one the second, and one said both contributed (though for the last person, the second was around a month ago). Not a great sample size, but seems like strong evidence that second-round long-covid isn't unheard of.

The one who said both contributed got their second case a month ago and already had severe issues the first time, so this is pretty much a sample size of one.

In addition to being less likely to happen at all (which definitely helps with Long Covid) later bouts are *much* less likely to be severe. If we have a model where severity is a major input, they're a lot less bad, as is Omicron.

This all makes me think of the problem of [trapped priors](#). If you think that long Covid is terrible all the time even in young healthy people who shrugged Covid off, you're going to think that later bouts must be terrible as well. If you think that people generally have a lot of health problems and blame them on Long Covid when they're primed to do so, you get the opposite conclusion from the same data.

Someone who is already super worried sees reports of bad problems after reinfections and thinks 'oh reinfections are bad too' whereas someone who thinks it's overblown thinks the opposite, that *this is evidence that the associations aren't causal*.

Both have a point. Conditional on it being real it got worse, but the probability of it generally being real should go down.

Katja's survey link is worth looking at if you want to dive deep and can adjust for the frame. It contains what seems like some good real data, although framed in a very non-neutral way. In particular I like that she asked about how much money people would give up in order to not have problems.

Her full sample size was n=57, these are 9 of the 11 reporting 'ongoing health problems.'

For people with ongoing health issues, given a choice of A) 'be rid of ongoing covid related health issues and symptoms forever' or B) an increase in income this year:

For 10% increase in income:

6 would take health, 3 income, 2 N/A

For 50% increase in income:

3 would still take health, 3 would take income

For 200% increase in income:

1 still prefers health, 2 would take income

I am not going to worry much about six people choosing health over a 10% increase in one year's income - it's 'problem goes away forever' versus a modest, one-time gain. Even 50% of one year's income is not that big a downside. In addition, one of the things impacted by health problems is income, including the impact to one's long-term career, so the health 'pays for itself' if it is at all a serious issue. None of this is a strong signal that we should be taking huge prevention costs.

I don't want to be the one person who would give up double their salary, no matter where the problems came from, especially since their willingness to pay might be much higher than that. I'm not so worried about the other eight.

O. It's not 100% that you will get it.

I'd guess there's a decent chance you'll be able avoid it, via reasonable-to-pay costs. For instance, maybe omicron basically brings herd immunity. Maybe rapid tests get more reliable, easy, and cheap. Maybe long covid becomes a bigger concern, and people worry enough to get R0 generally below 1. Or other technologies improve: Fend really does cut covid exhalations massively and is taken up widely, or something like Novid finds a way to be taken up. (Or your path takes you to being a hermit or living in China, or we spend next year hiding in bunkers from cyborg bees, and you are glad you didn't pay all that covid risk up front for parties that aren't happening.)

(If we were doing exactly what we are doing so far, but with air purifiers in most buildings, would R0 would have been $\frac{1}{4}$ instead of ~ 1 , and would it have died out? Is the problem that we are psychologically incapable of maintaining slightly more caution than needed to keep infections steady?)

Are things so predictable?

So this is where I see this kind of thinking as going 'off the rails.' Being a hermit or hiding in bunkers from cyborg bees? Sure, [living in China](#) is possible, although I doubt it will help that much as time rolls on, it's not like they can maintain zero Covid forever and if they do that's an additional reason not to live there even if you have the option under their policies. But the point here seems to be that you might go *completely insane* and then save on your *permanent* prevention costs? I mean, yeah, I guess, but that's very small comfort and impact on the math, and seems like one hell of a permanent cost.

I do expect things to quiet down a lot, but the idea that we might suppress this entirely is essentially impossible due to animals, and permanent prevention efforts are not worth it even if Long Covid is fully real. This is going endemic, and continuing to be paranoid when that happens is where this stops being remotely reasonable in my eyes.

I still support air filters and vaccinations on cost-benefit grounds, but we do not get to make collective decisions as individuals, and no we are not willing to maintain more caution than necessary *when levels are not actually risky* because why would we choose to do that.

I do agree you're not close to 100% to get it even if you do nothing at all, and never were. But that's one of the key arguments *against* prevention, that you don't go from 100% to 0%, with intense prevention retroactive efforts you go from 75% to 25% by the end of the Omicron wave or whatnot (numbers not intended to be very precise), which cuts the benefits in half or more. I still stand by Omicron largely not being worth preventing except during the few weeks when it was peaking.

This is also feeling very 'throw stuff at the wall.' There's worry about second infections and also hope that 'maybe you could never get a first infection' and that's part of an overall pattern that's looking to be terrified, although one could argue that at some point I started looking for ways to *not* be terrified and it's two sides of the same coin, so watch out for that.

P. The likelihood of you getting it probably does depend on how bad it is

This is a strange argument that I think goes the other way. If you're less likely to get Covid the worse it is, because in the worlds where Covid is more dangerous more people do more to prevent its spread, then that makes your forward-looking risk lower.

It also means that we can use the fact of people not caring much now as strong evidence that Long Covid isn't that big a deal. And that seems right to me.

Basically, if Long Covid was that bad, we'd all know about it through casual observation, it would be far more talked about and central in our culture, and there would be this *huge* pushback against people declaring the end of the pandemic. Since we don't see this almost at all among 'normal people' it means Long Covid is rare enough that normal people can mostly plausibly act like it's all fine. That implies low risk.

Q. Getting covid later is probably better than earlier.

So far this trend seems strong: I would rather get covid now than in March 2020. I expect more of this, from better knowledge, medicine, vaccines, and availability of everything we already have.

If I expect to get covid every year for the next five years, adding one more bout now is adding one more especially bad bout in expectation.

I agree with this as far as it goes, but already we've waited two years and already things are vastly improved, and this seems like an argument based on accepting that infection now *does* prevent a large fraction of an infection later if we hold the amount of prevention fixed. If it's a 'every year risk down by half' thing that's good news but the math this year is still the math this year.

There is one argument against this. It's better to get Covid while you're better protected, so if vaccines and boosters fade with time, delaying your infection (which then gives you even better protection) could backfire, if you don't intend to get continuous booster shots anyway.

Also, this has another implication that I agree with. We see a *lot* of the Long Covid cases people talk about coming from March 2020, or otherwise from very early. If Long Covid risk per case is already down quite a lot, we can put a reasonable cap on how bad it was in 2020 by looking at population-level graphs, then divide by a lot already. Yes, it would be better to be able to get Paxlovid than not, but we've already eliminated most of whatever risk we started with – at a minimum, >50% for Omicron, >50% for vaccination after that, and extra for better detection and treatment.

R. Huge VOI in following behind the crowd, at a safe distance

This is an argument that you don't want to get it *when everyone else is also getting it* and I did agree at the time that this was a thing, but I think its thingness has now ended. And I don't agree with the logic she lists here - I think it was wise because of hospital overload, rather than the benefit of waiting for a whole bunch of in-coming data from everyone else's cases. We have plenty of data.

Scott Alexander's Post

Scott's post has been mentioned a bit already, but there's a lot more there. His first section is about physical mechanisms, the second starts to get down to calculations.

2. The prevalence of Long COVID after a mild non-hospital-level case is probably somewhere around 20%, but some of this is pretty mild.

1. He then describes a bunch of studies: [Logue et al find](#) symptoms in 33% of patients vs. 5% of controls.
2. [The British Office of National Statistics](#) looks at people with a confirmed Covid test three months ago, and finds that 14% report having Long Covid symptoms, compared to 2% of a no-Covid control group.
3. [Haverfall et al in Sweden](#) found that 26% of people with previous non-hospital-grade Covid, and 9% of a control group, reported Long Covid-esque symptoms after 2 months. After 8 months, this was down to 15% and 3%.
4. [Sudre et al](#) got data from some kind of UK Covid app with four million users. They chose 4,000 who met various criteria and asked them about long Covid symptoms. 13% reported symptoms after a month, and 2% after three months.
5. [Thompson et al](#) get data from a UK longitudinal study. Their headline finding is that between 7.8% and 17% of patients seem to show at least one Long Covid symptom. But they have no control group, so probably it is lower than this. Also, only 1.2% to 4.8% of people say their Long Covid symptoms "impact normal functioning", which means a lot of people must have some annoying lingering symptoms that don't really bother them that much.

Scott's 20% seems to come from something like 'take a rough median of the excess reporting here' and to not worry about precision because there's so much vagueness built into the definition.

Remember that in my casual survey, about 32% of people reported having one of four symptoms *right now*, with almost zero correlation to previous Covid status.

When I saw the symptoms only showing up in 5% of controls in the Logue study, I thought if anything *that's* the weird one, so I looked and the control group is 21 'healthy individuals'. So that explains that.

There's no similar puzzle in study two because it asks about LongCovid specifically.

Third study only counts 'moderate to severe' reports in the above comparison, and it's a good thing it's of health care workers then because I have literally never met a civilian who ever reported having 'moderate to severe' anything that wasn't living in a pharmaceutical advertisement. This was their result in more detail.

Of the seropositive participants, 8% reported that their long-term symptoms moderately to markedly disrupted their work life, compared with 4% of the seronegative participants (RR, 1.8 [95% CI, 1.2-2.9]); 15% reported their long-term symptoms moderately to markedly disrupted their social life, compared with 6% of the seronegative participants (RR, 2.5 [95% CI, 1.8-3.6]); and 12% reported that their long-term symptoms moderately to markedly disrupted their home life, compared with 5% of the seronegative participants (RR, 2.3 [95% CI, 1.6-3.4]) ([Figure](#)). Furthermore, 11% of the seropositive

participants reported moderate to marked disruption in any Sheehan Disability Scale category as well as having at least 1 moderate to severe symptom lasting for at least 8 months, compared with 2% of the seronegative participants (RR, 4.5 [95% CI, 2.7-7.3]).

The differences in the different ratios here are interesting.

I will also repost this, [from a comment on a previous weekly post](#):

Via an excellent comment, we have an important discovery about the Long Covid data.

A major source for the previous pessimistic LC estimates, like Scott Alexanders (the UK's giant ONS survey) published an update of their previous report which looked at a follow-up over a longer time period. Basically they only counted an end to long Covid if there were two consecutive reports of no symptoms, and lots of their respondents had only one report of no symptoms before the study ended, not two, so got counted as persistent cases.

When they went back and updated their numbers, the overall results were substantially lower. This graphic explains their original mistake: The new headline result is 7.5% of Covid-19 patients had 'some limitation' of daily activities after 12 weeks if you ask them if they had long Covid-19. If you go by asking if there were any symptoms from a given list, the rate is lower (like 3%). The full report is here. What's notable is that a lot of participants reported LC symptoms with no Covid-19 positive test. They break it down by age and sex in the full report, but you should treat these numbers as numbers for mostly double vaxxed AZ and some mixture of single/double vaxxed Pfizer/Moderna for younger groups, since that's how it worked in the UK.

This is a pretty dumb error, a very dumb way to get a lot of people very scared and destroy a lot of value. Many thanks to the team for correcting the error, whether or not it was intentional and whether or not they should never have made the mistake. And whether or not the mistake was a reasonable one to be making, which it pretty much wasn't. Error correction is a big deal.

Basically what they did, as far as I can tell, was this: If you report symptoms, that means for now you have Long Covid. If you report no symptoms twice in a row, congratulations, you don't have long Covid. If you report no symptoms then symptoms, we still assume the symptoms are due to Covid, and you therefore still have Long Covid. If your last report was no symptoms, you're still considered to have Long Covid until you report again with no symptoms. A lot of people didn't feel the second no-symptom report was a terribly urgent thing to be doing. A lot of people simply hadn't had the chance to report a second time once their symptoms had cleared up. Yet they still counted the period that included their report of no symptoms, as a length of time that they had Long Covid.

To be blunt, they cheated (intentionally or otherwise), it was a massive effect, and we should have caught it, but to my knowledge none of us did. They have now fessed up. If you ask people to pick from a list of common symptoms, only 3% report that they have one. The larger numbers are mostly or entirely what happens when people are asked if there is anything wrong with them at all, and would they like to blame it on Covid-19. Also the percentages declined a lot over time, so chances are few of the cases would be permanent or semi-permanent. Even if you buy one of the larger numbers, this is a substantial improvement. Given how many people have already had Covid if you go by the antibodies present in various populations or what I would otherwise guess, this seems far more plausible, that Long Covid while real is relatively rare.

This kind of thing is *very* easy to miss, and I would never have found it without the comment.

3. The most common symptoms are breathing problems, issues with taste/smell, and fatigue + other cognitive problems.

And behold the mother of all COVID symptom persistence studies, [Amin-Chowdhury et al](#):

Table 2 Univariable analysis of symptoms by infection status (case, control group)

| Symptom group / symptom | Total N (%) N=1,300 | Case N (%) N=140 | Control group N (%) N=1,160 | P-value | Symptom group / symptom | Total N (%) N=1,300 | Case N (%) N=140 | Control group N (%) N=1,160 | P-value |
|--|---------------------------|------------------------|-----------------------------------|---------|---------------------------------------|---------------------------|------------------------|-----------------------------------|---------|
| Neurological | | | | | | | | | |
| Problems with sleeping through the night | 682 | 85 (60.7%) | 597 (51.5%) | 0.038 | Breathlessness after minimal exertion | 154 | 36 (25.7%) | 118 (10.2%) | <0.001 |
| Forgetfulness | 269 | 49 (35.0%) | 220 (19.0%) | <0.001 | Sore throat | 296 | 34 (24.3%) | 262 (22.6%) | 0.651 |
| Confusion/brain fog/trouble focusing attention | 209 | 39 (27.9%) | 170 (14.7%) | <0.001 | Chest tightness/pain | 121 | 26 (18.6%) | 95 (8.2%) | <0.001 |
| Tingling/numbness/needle pains in arms/legs | 168 | 25 (17.9%) | 143 (12.3%) | 0.065 | Fits of coughing | 94 | 19 (13.6%) | 75 (6.5%) | 0.002 |
| Seizures | 0 | 0 (0.0%) | 0 (0.0%) | - | Coughing when lying down | 80 | 14 (10.0%) | 66 (5.7%) | 0.045 |
| Collapse | 4 | 2 (1.4%) | 2 (0.2%) | 0.011 | Breathlessness at rest | 46 | 13 (9.3%) | 33 (2.8%) | <0.001 |
| Trembling | 28 | 8 (5.7%) | 20 (1.7%) | 0.002 | Asthmatic exacerbation | 51 | 8 (5.7%) | 43 (3.7%) | 0.248 |
| Twitching of fingers and toes | 36 | 8 (5.7%) | 28 (2.4%) | 0.025 | Having to sleep sitting upright | 21 | 4 (2.9%) | 17 (1.5%) | 0.217 |
| Short-term memory loss | 94 | 29 (20.7%) | 65 (5.6%) | <0.001 | Other (respiratory) | 12 | 3 (2.1%) | 9 (0.8%) | 0.110 |
| Trouble trying to form words | 129 | 22 (15.7%) | 107 (9.2%) | 0.015 | | | | | |
| Headache | 606 | 65 (46.4%) | 541 (46.6%) | 0.963 | | | | | |
| Hallucinations | 4 | 3 (2.1%) | 1 (0.1%) | <0.001 | | | | | |
| Dizziness | 169 | 25 (17.9%) | 144 (12.4%) | 0.070 | | | | | |
| Difficulty swallowing | 37 | 9 (6.4%) | 28 (2.4%) | 0.007 | | | | | |
| Other (neurological) | 46 | 19 (13.6%) | 27 (2.3%) | <0.001 | | | | | |
| Dermatological | | | | | | | | | |
| Dry/scaly skin | 355 | 38 (27.1%) | 317 (27.3%) | 0.963 | | | | | |
| Itchy skin | 343 | 38 (27.1%) | 305 (26.3%) | 0.829 | | | | | |
| Random bruising | 102 | 13 (9.3%) | 89 (7.7%) | 0.502 | | | | | |
| Rashes | 89 | 7 (5.0%) | 82 (7.1%) | 0.36 | | | | | |
| Hives | 41 | 6 (4.3%) | 35 (3.0%) | 0.417 | | | | | |
| Other (dermatological) | 22 | 2 (1.4%) | 20 (1.7%) | 0.798 | | | | | |
| Sensory | | | | | | | | | |
| Loss of smell | 35 | 26 (18.6%) | 9 (0.8%) | <0.001 | | | | | |
| Loss of appetite | 87 | 20 (14.3%) | 67 (5.8%) | <0.001 | | | | | |
| Loss of taste | 31 | 24 (17.1%) | 7 (0.6%) | <0.001 | | | | | |
| Ringing/stroke buzzing in the ears | 165 | 24 (17.1%) | 141 (12.2%) | 0.094 | | | | | |
| Blurred vision | 99 | 19 (13.6%) | 80 (6.9%) | 0.005 | | | | | |
| Metallic taste in the mouth | 64 | 14 (10.0%) | 50 (4.3%) | 0.003 | | | | | |
| Earache | 108 | 14 (10.0%) | 94 (8.1%) | 0.442 | | | | | |
| Sensitivity to light | 124 | 13 (9.3%) | 111 (9.6%) | 0.914 | | | | | |
| Flashing light in the eyes | 85 | 12 (8.6%) | 73 (6.3%) | 0.303 | | | | | |
| Pressure behind the eyes | 120 | 12 (8.6%) | 108 (9.3%) | 0.775 | | | | | |
| Slurred speech | 12 | 6 (4.3%) | 6 (0.5%) | <0.001 | | | | | |
| Respiratory | | | | | | | | | |
| Breathlessness after minimal exertion | 154 | 36 (25.7%) | 118 (10.2%) | <0.001 | | | | | |
| Sore throat | 296 | 34 (24.3%) | 262 (22.6%) | 0.651 | | | | | |
| Chest tightness/pain | 121 | 26 (18.6%) | 95 (8.2%) | <0.001 | | | | | |
| Fits of coughing | 94 | 19 (13.6%) | 75 (6.5%) | 0.002 | | | | | |
| Coughing when lying down | 80 | 14 (10.0%) | 66 (5.7%) | 0.045 | | | | | |
| Breathlessness at rest | 46 | 13 (9.3%) | 33 (2.8%) | <0.001 | | | | | |
| Asthmatic exacerbation | 51 | 8 (5.7%) | 43 (3.7%) | 0.248 | | | | | |
| Having to sleep sitting upright | 21 | 4 (2.9%) | 17 (1.5%) | 0.217 | | | | | |
| Other (respiratory) | 12 | 3 (2.1%) | 9 (0.8%) | 0.110 | | | | | |
| Gastrointestinal | | | | | | | | | |
| Bloating | 331 | 38 (27.1%) | 293 (25.3%) | 0.629 | | | | | |
| Nausea | 197 | 29 (20.7%) | 168 (14.5%) | 0.052 | | | | | |
| Abdominal cramps | 224 | 27 (19.3%) | 197 (17.0%) | 0.496 | | | | | |
| Diarrhoea | 244 | 26 (18.6%) | 218 (18.8%) | 0.949 | | | | | |
| Constipation | 200 | 20 (14.3%) | 180 (15.5%) | 0.703 | | | | | |
| Vomiting | 34 | 6 (4.3%) | 28 (2.4%) | 0.190 | | | | | |
| Fecal incontinence | 17 | 4 (2.9%) | 13 (1.1%) | 0.088 | | | | | |
| Other (gastrointestinal) | 8 | 0 (0.0%) | 8 (0.7%) | 0.324 | | | | | |
| Cardiovascular | | | | | | | | | |
| Cold hands/feet | 285 | 32 (22.9%) | 253 (21.8%) | 0.777 | | | | | |
| Palpitations | 163 | 30 (21.4%) | 133 (11.5%) | 0.001 | | | | | |
| Low blood pressure | 67 | 11 (7.9%) | 56 (4.8%) | 0.126 | | | | | |
| High blood pressure | 60 | 7 (5.0%) | 53 (4.6%) | 0.818 | | | | | |
| Excessive bleeding from cuts/wounds | 17 | 2 (1.4%) | 15 (1.3%) | 0.894 | | | | | |
| Other (cardiovascular) | 6 | 2 (1.4%) | 4 (0.3%) | 0.074 | | | | | |
| Blood clots | 2 | 0 (0.0%) | 2 (0.2%) | 0.623 | | | | | |
| Mental health | | | | | | | | | |
| Stress | 742 | 76 (54.3%) | 666 (57.4%) | 0.48 | | | | | |
| Anxiety | 619 | 72 (51.4%) | 547 (47.2%) | 0.339 | | | | | |
| Difficulty to sleep at night | 566 | 69 (49.3%) | 497 (42.8%) | 0.147 | | | | | |
| Frustration | 581 | 60 (42.9%) | 521 (44.9%) | 0.644 | | | | | |
| Sadness | 543 | 59 (42.1%) | 484 (41.7%) | 0.924 | | | | | |
| Difficult to wake up in the morning | 436 | 47 (33.6%) | 389 (33.5%) | 0.993 | | | | | |
| Mood swings | 436 | 54 (36.6%) | 387 (32.9%) | 0.182 | | | | | |
| Depression | 277 | 31 (22.1%) | 246 (21.2%) | 0.798 | | | | | |
| Loneliness | 302 | 24 (17.1%) | 278 (24.0%) | 0.071 | | | | | |
| Nervous breakdown | 39 | 4 (2.9%) | 35 (3.0%) | 0.916 | | | | | |
| Other (mental health) | 13 | 3 (2.1%) | 10 (0.9%) | 0.150 | | | | | |
| Other | | | | | | | | | |
| Unusual fatigue/tiredness after exertion | 258 | 55 (39.3%) | 203 (17.5%) | <0.001 | | | | | |
| Muscle aches/pains | 357 | 43 (30.7%) | 314 (27.1%) | 0.361 | | | | | |
| Neck/shoulder pain | 467 | 41 (29.3%) | 426 (36.7%) | 0.083 | | | | | |
| Joint pains | 328 | 40 (28.6%) | 288 (24.8%) | 0.335 | | | | | |
| Back pain | 443 | 40 (28.6%) | 403 (34.7%) | 0.146 | | | | | |
| Disruption of the menstrual cycle | 110 | 17 (12.1%) | 93 (8.0%) | 0.098 | | | | | |
| Sinus pain | 129 | 14 (10.0%) | 115 (9.9%) | 0.974 | | | | | |
| Hair loss | 99 | 13 (9.3%) | 86 (7.4%) | 0.430 | | | | | |
| Lymph nodes/glands (parotid/neck/ears, etc) | 81 | 12 (8.6%) | 69 (5.9%) | 0.225 | | | | | |

AC&E act as if this is reassuring – their conclusion starts with “most persistent symptoms reported following mild COVID-19 were equally common in cases and controls” – but it really isn’t. Not only does this 8-month-out sample find high levels of the expected problems (fatigue/smell/taste/breathing), but it finds some unexpected ones too. Cases are likelier than controls to have cognitive problems and weird neurological issues. One flaw in this analysis is that it didn’t ask for premorbid functioning, so you can tell a story where unhealthy people are more likely to get COVID than healthy ones (maybe they’re stuck in crowded care homes? Maybe they put less effort into staying healthy in general?) But I don’t think this story is true – how come obviously plausibly COVID linked things (like smell problems) are significant, and obviously-not-COVID-linked things like diarrhea aren’t?

One thing this study *does* reassure me about is mental health. A lot of people claim that long COVID involves various mental health sequelae. This study comes out pretty strongly against it. Sure, lots of COVID patients are depressed – but so are equally many controls. The age of COVID is just a depressing time. In fact, it’s kind of weird that you can get this much fatigue, brain fog, etc without an increase in depression diagnoses.

The lack of anything happening in the mental health category jumped out to me as well here. If there’s this many bad symptoms running around and a lot of people have fatigue and brain fog then it’s more than *kind of* weird not to have depression and loneliness and

sadness and frustration follow in their wake. I can tell you right now, if I had severe brain fog, *I strongly predict I would be highly frustrated.*

4. Sometimes problems go away after a few months, other times they don't

I mean, yes. Thanks. You can read his descriptions but as far as I can tell this is mostly the throwing up of hands because no one knows. The scary part about chronic fatigue is long-existing chronic fatigue studies in general, not Covid-related.

5. Psychosomatic symptoms probably aren't the majority of long COVID.

I mean, I'm not seeing too many people claiming that they are. There are a lot more people worried that someone else might be claiming that, than people actually making the claim.

He then reminds us that chronic fatigue syndrome is really bad. I will affirm that I believe chronic fatigue syndrome is really bad, I guess?

The claim I'm making earlier is not that people are 'making all this up' or anything, it's that they are often attributing it to Covid when it's there for other reasons. That doesn't make their lives suck less. Nor would it being psychosomatic. Again, still makes life suck. I do think that the French study suggests that something like half and perhaps more of the Long Covid claims are being misattributed.

6. Long COVID is probably rare in children

My overall conclusion here is that long COVID is rarer in children than adults, and may not exist at all. The studies tell us it's probably somewhere less than 5% of kids, but so far we can't conclude anything stronger than that.

Note that this is in the model where Scott is defining 'Long Covid' so broadly that 20%+ of Covid cases count, so this is a 75%+ reduction. I agree that the evidence here points in the direction of Long Covid being rare in children, and of course that counts as more strong evidence in favor of severity impacting the chances of getting Long Covid, and also some of the details show the dangers of the misattribution problem.

7. Vaccination probably doesn't change the per-symptomatic-case risk of Long COVID much

He links to [this Twitter thread](#) that puts the long Covid rate at 1.04% among breakthrough symptomatic cases, based on self-reports - I worry this link was put in wrong because Scott says it's complicated but it doesn't seem like a complicated thread. I'd also note this:

(all this information is from an online poll by a sketchy group of COVID "survivor" activists. But they wrote up their poll in [the scientific paper font](#), as a PDF and everything, so I say we count it anyway)

I can't tell if he's joking, not sure if Scott can tell either. I need to use the power of the scientific paper font.

Whereas the next link, [this NEJM study](#), found 19% Long Covid after six weeks by asking about the usual grab bag of symptoms.

Then he says this:

And just before publishing this, someone sent me [this study](#), which very preliminarily finds vaccines might decrease Long COVID risk by a factor of 2. I think a factor of 2-3 is believable; one of 10 or 20, less so.

This seems right to me if we're talking conditional on symptomatic infection (and bonus points for the title 'short report on long Covid.') So I'll pencil in a factor of two for vaccination conditional on *symptomatic* infection, in addition to the reduction in risk of such infections in the first place, and in addition to another factor of two (perhaps more) for Omicron.

And finally this:

Weirdly, there are some claims that [vaccines can help relieve symptoms of existing long COVID](#). Sounds kind of like sympathetic magic to me, but the researcher quoted in the linked article said it might "improve symptoms by eliminating any virus or viral remnants left in the body" or by "rebalancing the immune system". So yeah, sympathetic magic.

I mean worth a shot, I guess?

At one point I linked to [this Twitter thread](#) citing 50%-80% reduction in Long Covid symptoms in vaccinated individuals.

8. Your risk of a terrible long COVID outcome conditional on COVID is probably between a few tenths of a percent and a few percent.

This is the big one so I'm going to quote the whole chain of logic.

My original calculation went like this:

About 25% of people who get COVID report long COVID symptoms. About half of those go away after a few months, so 12.5% get persistent symptoms. Suppose that half of those cases (totally made-up number) are very mild and not worth worrying about. Then 6.25% of people who get COVID would have serious long-lasting Long COVID symptoms.

After doing that calculation, I read [this essay](#) by Matt Bell, who tries to figure out the same thing. He is much more optimistic. He agrees that about half of long COVID cases go away after a few months, but adds another 50% decrease from "few months" to "lifelong", kind of on priors, admitting there's not too much positive evidence for this. Then he adds another factor-of-two decrease from vaccination, based on [very preliminary studies from the UK](#). He estimates that someone with my demographics (vaccinated man in his 30s) has a 2% risk of Long COVID conditional on getting COVID at all. Then he divides by five for the true worst case scenario, based on studies showing that a fifth of people with Long COVID report that it affects their daily activities "a lot". So by his final number, I have an 0.4% chance of getting really terrible long COVID, conditional on getting COVID at all.

My friend AcesoUnderGlass also did [a writeup of this](#), published after I did my first-draft calculation, which seems to be thinking of this very differently, based entirely on hospitalization rates (which of course are very low in vaccinated people our age). She accordingly concludes that risk is very low. I don't really understand her reasoning here, but I trust her a lot and am working on trying to converge with her on this.

This sounds a lot like a punt, where the evidence is murky and Scott is simply giving error bounds. Error bounds do seem useful, so let's see what we can say.

For the (reasonable) upper bound I do think we can start with a 25% chance of 'any long Covid symptoms at all' based on symptomatic Covid infection that took place early in the pandemic. This is intentionally a worst-case scenario. Then we can start dividing, and Scott's logic seems reasonable enough. I can't imagine less than half of people's symptoms fading given a reasonable amount of time, and we know that such symptoms only cross this kind of threshold a minority of the time. That would get us to Scott's 6.25%.

However, that's *symptomatic* Covid, and that's without vaccination or improved treatments, which I think are safely another two divisions by two. And we can add another adjustment for

Omicron since almost all forward-looking cases are Omicron, which divides by two again. It's also without an adjustment for one's previous health, or a number of other things, but this is an upper bound.

So the pessimistic upper bound would be:

1. 25% base case of anything at all if you're symptomatic under base conditions
2. 50% reduction from vaccination
3. 50% initial chance of being symptomatic pre-Omicron pre-vaccine
4. 50% reduction from Omicron
5. 50% reduction for chance it doesn't stick around for long-term
6. 50% reduction for whether 'it's as bad as all that.'

Take all of that together and we get a number around 0.8% as a *reasonable upper bound*.

Another way to get a (reasonable) upper bound is to look at the population-level statistics and ask what we would be noticing on that level, or what matches our observations within our networks, even vaguely, and seems similar.

What should be the (reasonable) lower bound? Very low.

If I wanted to use the above method and keep dividing, I'd adjust the initial 25% to 20% or so, assume that vaccinations cut risk by a factor of 3 rather than 2 especially when we include a booster, assume that two thirds of cases are asymptomatic rather than half even before Omicron, give Omicron credit for a 75% reduction instead of 50% (especially given less reports of loss of taste and smell), and say that half of what we've observed is misattribution or needlessly psychosomatic (as in, if you don't believe in Long Covid it won't happen in those cases). Then we'd want to adjust for your health, which could easily be another 50%+ reduction for good initial health. And we can cut the 'it's as bad as all that' down from 50% to 20% for the same reason Matt does, and add a 50% reduction for things sticking around forever rather than for the medium-term like Matt does.

So the optimistic-but-not-crazy lower bound would be:

1. 20% base chance of anything at all if you're symptomatic under base conditions.
2. 66% reduction from vaccination
3. 66% initial chance of being symptomatic pre-Omicron pre-vaccine
4. 75% reduction from Omicron
5. 50% reduction for chance it doesn't stick around for medium-term
6. 50% reduction for chance it doesn't stick around for long-term
7. 80% reduction for whether it's 'as bad as all that.'
8. 50% reduction for misattribution or unnecessary psychosomatic cases
9. 50% reduction for good initial health and selection impacts

That multiplies to a 0.007% chance, which you'd *then* have to multiply by the chance you can actually prevent a case moving forward, and which is already well below the 'ignore this' threshold. I could find additional excuses to keep going, but the risk of double-counting is a thing and the point has been made.

If I went with conservative numbers as a precautionary principle estimate:

1. 25% chance for literal anything at all, sure, why not.
2. 66% reduction from vaccination plus booster (but requires booster)
3. 50% reduction from being asymptomatic early on and then this went up later, combined with asymptomatic cases not being quite totally safe
4. 50% reduction from Omicron only, on precautionary principle
5. 50% reduction for sticking around
6. 0% reduction from the second sticking around reduction because I'm worried about accidentally double-counting somewhere and want to be safe

7. 60% reduction for 'as bad as all that' based among other things on my survey
8. 25% reduction for misattribution to give benefit of the doubt
9. 0% impact of good health

That gets us back to a 0.2% chance of Long Covid conditional on first infection, and less than that less for all future ones combined because of immune strengthening. That's *still* the biggest consideration in many cases on whether to care about preventing Covid, but your life is ending one minute at a time.

We can then double back and look at the anecdata and population-level knowledge we have to compare. A 0.2% forward-looking chance for a boosted, essentially healthy person of 'some serious s***' seems fully consistent with my observations.

My guess is that the real number is lower, but given uncertainty and precaution I think it's fair to treat it as about 0.2%.

We then can multiply by the chance of getting Covid in a given year going forward, to see one's per-year risk. Or we can multiply by the chance of *preventing* Covid in a given year via ongoing prevention measures you're considering, and multiply by that to get the *benefit of prevention*.

Here are Scott's conclusions:

10. Conclusions

1. Long COVID is many different issues without a common mechanism.
2. Some of these are straightforward and not surprising, eg lung scarring and post-ICU syndrome from severe infection, and would happen in any disease of this severity. Others seem to be more like the poorly-understood postviral syndromes associated with several other diseases. While some symptoms may be psychosomatic, most are probably organic.
3. The three major categories of symptoms are straightforward cardiovascular-pulmonary issues, straightforward smell and taste issues, and more mysterious neurological issues.

Worth noting that Omicron seems to have a lot less issues with loss of taste and smell.

4. Although these get better with time in some people, in a significant number (maybe ~50% of people who had them at six weeks) they persist for as long as anyone has been able to measure them (a few months in the case of COVID, a year or two in the case of comparable syndromes).
5. Post-COVID fatigue is particularly concerning. This would be very bad if we analogized it to CFS/ME, and still pretty bad if we analogized it to other known postviral syndromes. There is no proof that this always gets better over the long term, although no study has looked at them for more than a few years. Facing postviral fatigue on this scale is a new problem.
6. Children probably get Long COVID less than adults, probably at a rate of less than 5% of symptomatic cases. But we don't know how much less, and we can't rule out that some children get pretty severe symptoms.
7. Although vaccination decreases the risk of symptomatic COVID, it probably doesn't decrease the risk of Long COVID per symptomatic COVID case by very much, though it might decrease it by a factor of 2-3.
8. Your chance of really bad debilitating lifelong Long COVID, conditional on getting COVID, is probably somewhere between a few tenths of a percent, and a few percent.

Your chance per year of getting it by living a normal lifestyle depends on what you consider a normal lifestyle and on the future course of the pandemic. For me, under reasonable assumptions, it's probably well below one percent.

EDIT: Here are some other people who tried to do this same analysis. I learned about all of these after I wrote the first draft of this, so you can consider the basic thought process here to be independent of them – but I edited some things to account for what I learned from them before writing the final version.

- AcesoUnderGlass: [Long COVID Is Not Necessarily Your Biggest Problem](#)
- Matt Bell: [If You're Vaccinated, Your Main Risk From The Delta Variant Is Probably Long-Haul COVID](#)
- 1DaySooner: [FAQ: Long-Term Effects of COVID-19](#)

Other Things Of Note

I searched for the term 'Long Covid' over all of my weekly posts, to see if I was missing anything important that I'd said previously. Here are the passages that didn't fit in above but still seem relevant, including to the mindset.

From 9/17/21:

[This the latest effort to quantify Long Covid](#), which I found a link to via LessWrong, with the author thinking of it as similar to chronic fatigue syndrome. There's enough thoughtful effort here that it needs to be included, but what sticks out at me is the continued reference point shifts in what constitutes Long Covid.

Surveys ask people if they have any symptoms at all, that's considered Long Covid, then that's compared to CFS because the most typical symptoms are most typical of CFS. Therefore create association of Long Covid as equivalent to lifelong crippling fatigue. And as usual, I have zero faith in what passes for controls in such measurements. There are then some calculations I'd argue with even given the premise. And if things really were this bad, and a good fraction of a percent of people who got Covid ended up permanently crippled, many of them unable to work normal jobs, I'd point out how many people already did have Covid by this point, it's at least a large minority of the population. So this seems like the kind of thing that would be impossible to miss on a population level.

On [this Washington Post article](#), 11/18/21:

Then the CDC comes out and says that a full fifty percent of people get Long Covid, with symptoms that persist for at least six months.

If that was actually true, we would know, because a large percentage of the population got Covid-19, so the population statistics would be screaming at us.

There's this great tidbit: So you're telling me the percentage is the same at one month as it is at six months, and also two to five months?

This has got to be some sort of fishing expedition for fishing expeditions. My combined update from this week is to move closer to the French study's position.

[The German study](#), on 12/30/31:

New paper out of Germany. Dashboard link.

I'm surprised that 61.9% of positives had already been caught by PCR. Finally, a control group, and it looks like it controlled for quite a bit. Assuming this is being reported

correctly, it's saying that if you have an unknown infection, you don't get meaningful Long Covid.

This at least rules out the hypothesis that Long Covid is a threat in asymptomatic cases. If you don't realize you had Covid, your rate of reporting symptoms doesn't rise, at least not a substantial amount. We could in theory give the benefit of the doubt that everyone with symptoms got a PCR test, but I'm still awfully suspicious of zero effect, especially when other claims sometimes involved not differentiating much based on severity of the infection.

To the extent that Long Covid is a non-placebo Actual Thing, this seems to strongly suggest that it will indeed scale with the severity of infection, so vaccinations and booster shots will help a lot, and if you're at low risk for Short Covid (as it were), you're also therefore not at such high risk for Long Covid either. The idea that children were getting permanently crippled by otherwise harmless infections all over the place doesn't match this data at all. Long Covid, and the rise in morbidity in the year after Covid infection, are likely the result of Covid infections that get bad and do a bunch of damage.

That also implies that Paxlovid and other early treatments will improve Long Covid outlooks.

There's a long and good analysis in [the Long Covid section of this post](#) from 07/29/21, but it uses a lot of pictures and is long so I'm not copying it over.

The biggest continuing theme is that a *lot* of times I have had to address scaremongering about Long Covid in children, who are at very little risk.

Conclusion

I did call this The Long Long Covid post for a reason. I put off writing this for too long, but hopefully it will still be of some use. Going forward, I strongly believe that while Long Covid is real and scary and the main thing to worry about in terms of what remains of the pandemic, it is not scary enough that it should cause us to fail to live our lives. The stakes simply aren't high enough for that, and in particular children are at minimal risk.

I already put my main conclusions up top because of tl;dr, so here I'll end with an even shorter practical summary.

1. Children are at minimal risk.
2. Risk from a given case is proportional to its severity.
3. One can act as if serious Long Covid will occur in ~0.2% of boosted Covid cases.
4. That's not nothing, but it's not enough that you shouldn't live your life.

Patient Observation

Knowing the territory takes patient and direct observation.

When a person first begins to study naturalism with me, I say to them almost nothing that I've written in this sequence.

That might very well be a mistake; I might get better results if I knew some short combination of words that would cause them to correctly understand, intellectually, what we are doing and why, before we started.

But in fact I bank on the person's trust in me a bit, and begin by helping them establish consistent habits of observation.

If they're interested in studying confusion, I ask them to tap their leg every time they notice they're confused. I ask them to keep a log book in which they record a few words about their experiences of confusion each day. I ask them to make predictions about what it will be like to notice confusion—what kind of situation will be happening and how they will know in the moment that they are confused—and to compare their observations to their predictions.

And then, throughout what has so far proven to be about a three month program, I never shift our focus away from consistent habits of observation. It's not just where I start. It's the entire curriculum.

From a practical perspective, this dogged persistence is the foundation of naturalism. "Direct observation of the territory", without patience, gets you something like a bag of tricks. Valuable tricks, but still tricks. Isolated mental motions made when they are convenient and enjoyable, not when they are most needed.

With patience, though, you get a life-long practice of epistemic rationality.

So the whole naturalism program, from start to finish, consists of the establishment, improvement, and maintenance of consistent habits of observation. Consistent and ceaseless, without any sense that "and then we'll be done observing and get back to normal". When my students and I meet, we are constantly talking about what daily practice looked like over the past week, what was too heavy to implement as a regular routine, and how they personally can slow down enough to thoroughly observe. We continue in this way until they no longer seem to need me to ask those questions for them. The program is meant to introduce a new normal.

The rest of what I talk about in this sequence is sprinkled in, sure. It's implicit in the questions I ask, the approaches I encourage, and "my whole vibe". Like a catalyst, or like spices, those other parts are crucial to the recipe. But patience is the engine that makes all of it go. None of what I've written here gets off the ground unless it is practiced as an ongoing discipline.

I suspect that the thing I'm calling "patience" really is a single core capacity, or a single virtue; but it can express itself in multiple patterns of behavior. I'd like to talk about three patient behavior patterns: **tenacity, openness, and thoroughness**.

In my mind, the paradigmatic example of **tenacity** is marathon training. Not marathon racing, but the training program a runner goes through to prepare for their first marathon.

After an initial adjustment period, training for endurance athletics is mainly difficult because it requires commitment to the maintenance of a routine, not because it requires intense exertion. Most of the time you're not running anywhere near as fast as you can, and at least until race day, you're not running as far as you can either. On any given day, you're running a comfortable-for-you amount—but you're doing it day after day after day, without fail, for months at a time. The distance running motto is “small, consistent efforts”.

The facet of patience I'm calling tenacity is the ability to exert small, consistent efforts.

The reason tenacity is foundational to naturalism is that it's required for any kind of maintenance. Knowledge of the territory requires not just contact with the territory, but *maintenance* of that contact. Bumping up against the world and bouncing right back off again is not enough; you have to reliably *return* after you bounce.

We are bound to see things as we are, rather than as *they* are; but we are not bound to always see them as we were when we first encountered them. It is possible to observe again, and again, and again; if you do that with naked directness, and with the relentlessness of marathon training, your perceptual systems will inevitably adjust to perceive reality more accurately over time.



The next facet of patience I'd like to talk about is “**openness**”, in the sense of “non-closure”.

When someone comes to me for advice on a long-standing adaptive challenge, the most common recommendation I give is, “Stop trying to solve this problem for a while. Start investigating the underlying territory instead.”

I recently chatted with someone who was worried that she might be a narcissist, and wanted to know what to do about it. She gave me permission to share these (anonymized) excerpts from our conversation.

Logan: my first thought is that this sounds like a situation where you'd do well to put "what should i do about it?" on hold for a good three months, and focus instead on "what is actually happening? how can i tell? what is it like? what is my brain doing by default in various situations, and which situations are the ones i care about here? which phenomena and mental motions seem important for understanding what's happening here?"

Crystal: That seems smart. But, seems better to know I am a narcissist than to be uncertain about it for three months... more comfortable I mean

Logan: i imagine that you have a question like "am i a narcissist?" in your head, and it's really salient because things you care about depend on the answer to it, and it's uncomfortable to not know the answer because you'd ideally orient to the two different worlds differently, and when you don't know which you're in you don't know how to orient. is that right?

Crystal: Yeah

Here, Crystal is demonstrating a *need for closure*. She is uncomfortable (understandably!) with being uncertain. She would like to make plans for the future. Those plans may be substantially different in worlds where she believes the answer to the question "am I a narcissist?" is "yes" than in the world where it's "no." She wants to know what to work on, in herself and in her relationships. She wants to know what to expect.

So when I recommend to her that she deliberately *hang out* in uncertainty while she gradually increases her contact with the territory, it feels bad to her.

Back to the conversation, jumping ahead a bit:

Logan: i tend to operate under the conjecture that when there is a thing that's been a problem for most of a person's life, that person's way of conceptualizing the problem is very likely to be incorrect or incomplete in ways that make investigation that's not driven by the concept more productive than investigation that *is* driven by the concept.

Crystal: The concept being "narcissism"?

Logan: yeah.

Crystal: How is it driven by the concept or not? What does that mean?

Logan: investigation that's driven by the concept looks like: how would i know if i'm a narcissist? what things are evidence for or against? where would i look for evidence of narcissism? what would disconfirming evidence look like? what are alternative hypotheses? how might i test them?

Crystal: > investigation that's driven by the concept looks like: how would i know ...

That sounds like me

Logan: investigation that's *not* driven by the concept might look like: what does it feel like to be worried about whether i'm a narcissist? what seems to be at stake? if i go through the week and write down times when something related to the-thing-i-care-about-here happened, what do i end up writing? what was happening around me and in my head during those times? which of the things happening in my experience seems most closely tied to the-thing-i-care-about-here? how can i tell when that thing is happening in my head? if i watch for times when that thing is happening in my head and write down instances, what do i write?

in other words, it's possible to gain a lot of information about what's actually going on without having pre-decided most of what's going on. my suspicion is that this is a time when it makes sense to not pre-decide most of what's going on before you try to really seriously get in contact with the relevant region of territory

I often call the latter type of investigation—the kind that's *not* driven by the concept—"exploratory investigation". I've never used a word for the former type, but here I'm inclined to dub it "certainty seeking".

Certainty seeking is often the right approach. It's the right approach when you have good reason to think you mostly understand the situation and just need to fill in some details, or to determine the truth values of a couple central propositions. In that case, a more exploratory investigation style would be needlessly inefficient.

But people very often fall into certainty seeking when they are *impatient*. They already have a sketch, and for one reason or another, they just want to fill in the details and be done. They're willing to shift a line here or there, but mainly they're motivated to *complete the drawing*. "All I want to know is, is this narcissism or isn't it? Yes or no?!"

A person in the grips of this impatient mode is not so much trying to learn the shape of reality, as to crystalize a satisfying concept so they can relax into certainty.

There are advantages and disadvantages to both approaches, of course. My point is that nearly all truth-seeking benefits from a *combined* approach to investigation. You need to be able to move back and forth. Impatience tends to crowd out direct exploration, and ensures that you'll mainly find whatever you have already decided to look for.

Openness, in the sense I mean, is the ability to observe *without desperation for an answer*.

But, what is it to observe without desperation? I've told you what this facet of patience *lacks*, but what does it *consist of*?

Today I saw a raven do a barrel roll^[1].

I'd heard that ravens could turn over in the air, and even do backflips occasionally, and I'd seen pictures of ravens upside down. But when I saw this one do a barrel roll right in front of me—well, above me, I suppose—I felt... "surprise" is too simple. I felt glued to the ground, knocked sideways, and opened up all at once. I felt awe. I shouted up to the raven, "You just did a barrel roll! What?! That was awesome! You are awesome!"

Before I saw the raven, I was out on a walk through the country, down a dirt road with a few houses and lots of trees. Earlier on my walk, I took a picture of some kind of insect nest, or perhaps a fungus, on the underside of a leaf. I peered through a hedge to see if I could work out what kind of crop my secretive neighbors were growing. I smelled some little pink flowers on a tree and found that their scent was a lot like caramel and roses mixed with grass. I pet a dog and asked her if she could smell my cat (which she probably could, but she wasn't feeling chatty). I learned that the acorn hats have dried out enough to go "crunch" underfoot.

My state of mind was one of open, gentle exploration. And it's from that state of mind that the raven was able to move me in the way it did.

I can imagine an alternate walk in which I was trying to determine whether or not my local ravens can do aerial acrobatics. I think there would have been some frustration with the many ravens I saw along the way who were not even flying, let alone flying upside down. (I wouldn't have observed any leaf bottoms at all.) And I think that seeing the barrel roll would still have been very cool, but it also would have felt a lot more like relief and completion; like the end of something, rather than the beginning. Like closing, more than like opening.

But more importantly, I never would have set out on such a walk in the first place. It simply would never have occurred to me. I saw a raven do a barrel roll because I was *there* when it happened. I was in the right place, and I was open.

Openness feels like being there for whatever happens. Being *down*.

It's almost-but-not-quite the opposite of purposefulness. It's the canvas on which purposes get painted. A central strategy of naturalism is to put most of your purposefulness points into choosing where to bring your canvas. If that canvas is already full, then there's nowhere for new and surprising things to land.

Openness feels like putting myself in the middle of something *alive*, looking around, and letting whatever I observe move me however it does.

The final facet of patience I'd like to discuss is **thoroughness**.

When I think of thoroughness, I imagine holding a puzzle box as I turn it around and around, trying to visually examine it from all angles. No matter how accurate and precise my observations of the box from one particular angle, it is only possible to see at most three faces of a cube from any single vantage point. To know the whole surface of a cube, I either have to move the cube, or I have to move myself.

If the puzzle box is sitting on your desk, and you glance over at it several times as you go about your day, you'll most likely catch it from a few different angles by accident. So tenacity and openness together naturally result in some amount of thoroughness.

This is the principle behind what Anna Salamon has called "the 50/50 rule". According to (my own interpretation of) the 50/50 rule, 50% of the intellectual progress you make on something will happen while you are deliberately trying to make progress on that thing in particular. The other 50% will happen while you are engaged with other things: riding the bus, playing with your kids, designing a board game, identifying a bird.

It's important to spend a lot of your time doing things *other* than focusing on your Main Project. This is not *just* because your brain needs to "rest"; it is also important to do other things because you will see different faces of the puzzle box while you are dancing at a salsa club than while you are staring at a white board.

It is possible to find additional vantage points on purpose, and I call this capacity "perceptual dexterity".

When I look at the pen that is on my desk right now, I see a pen. That is, when I direct my gaze and attention toward the part of my visual field where light is reflecting off the surface of the pen, my concept of "pen" is active.

When I see the pen as a pen, certain parts of my experience stand out to me, while others are discarded. My attention lands on the button at the top, which I could push to extrude the nib. It lands on the thin cylindrical shape, and I can feel myself preparing to orient my hand to that shape in a way that would allow me to hold the pen for writing.

But the reflection of the clip in the shiny metal surface of the cylinder *doesn't* occur to me, when I see the pen as a pen. To notice that reflection, I have to see the pen a little differently than I would by default. I have to rotate to a slightly different vantage point—to rotate my *mind* into a slightly different configuration, one that processes information a bit differently.

I don't have to primarily activate my "pen" concept just because I happen to be looking at a pen. I can choose to rotate my mind however I want, and *then* look at the thing in front of me.

If I rotate my mind toward “goose”, what first stands out to me is the hole in the front, which seems to break an otherwise aerodynamic nose. Maybe the air would get stuck in there, if this pen had wings and tried to fly.

If I rotate toward “aggression”, the first thing that stands out is the place where the clip is attached to the body of the pen, as I evaluate its thickness and wonder how much force it would take to snap the clip off and leave a sharp edge.

When I’m chewing on a problem and feeling a little stuck, one of the first things I do is ask myself, “If this were a boat, what sort of boat would it be?” There’s nothing special about boats for problem solving, but answering this question forces me to rotate my mind into a configuration that is probably quite different from whatever I was stuck in before. If this pen were a boat, it would be a sleek but sturdy racing boat with a silver sail and an athletic captain steering.

The more perceptually dexterous you are, the less constrained you are to see only what you saw in your very first glance. You are not trapped in your most familiar perspective.

Thoroughness is what results from the successful exercise of perceptual dexterity. It is observation that continues well beyond familiarity, traversing many vantage points to triangulate reality.

The thing that tenacity, openness, and thoroughness have in common is what I mean by “patience”. It’s the opposite of “rushing”, the opposite of “just wanting to be done”, or the opposite of “jumping to conclusions”. Patience is taking the time to discover the real shape of the world.

By “patient observation”, then, I mean observation that is tenacious, open, and thorough. Knowing the territory requires direct observation that *takes its time* to discover the shape of the world.

1. ^

It was technically an aileron roll.

Prizes for the 2020 Review

The third annual review is complete. We spent two months [nominating](#), [reviewing](#), and [voting](#) on the best posts of 2020. After lots of thoughtful evaluation, it's time to award some prizes!

As described in the [voting results post](#), there are two prize pools for posts. Lightcone Infrastructure contributed \$10k to the Review Vote prize pool, and LessWrong users donated \$1770 to the Unit of Caring prize pool (which Lightcone matched 1:1, bringing it up to \$3540).

Lightcone Infrastructure is also awarding \$3600 in prizes for people who wrote reviews^[1] for posts.

Prizes for Posts

The Review Vote Prizes

We've awarded a \$1000 Top Prize to each of the following authors:

- **catherio**, for [microCOVID.org, a tool to estimate covid risk](#)
- **Ajeya Cotra**, for [Draft report on AI Timelines](#)
- **evhub**, for [An Overview of 11 Proposals for Building Safe Advanced AI](#)
- **johnswentworth** for [When Money is Abundant, Knowledge is the Real Wealth, Alignment by Default, The Pointers Problem, and Coordination as Scarce Resource](#)

We've awarded a \$500 Honorable Mention to each of:

- **Alkjash**, for [Pain is not the unit of effort](#)
- **Mark Xu**, for [The Solomonoff Prior is Malign](#)
- **Jacob Falkovich**, for [Seeing the Smoke](#)
- **Alex Flint**, for [The Ground of Optimization](#)
- **Zvi**, for [Simulacra Levels and their Interactions](#)
- **Richard Ngo**, for [AGI Safety from First Principles](#)
- **Paul Christiano**, for [Inaccessible Information](#)
- **Steven Byrnes** for [My Computational Framework for the Brain](#)
- **abramdemski** for [Most Prisoner's Dilemmas are Stag Hunts; Most Stag Hunts are Schelling Problems](#), and [An Orthodox Case Against Utility Functions](#)
- **Daniel Kokotajlo**, for [Cortés, Pizarro, and Afonso as Precedents for Takeover](#)
- **Scott Garrabrant**, for [Introduction to Cartesian Frames](#)
- **Rafael Harth**, for [Inner Alignment, Explain Like I'm 12 Edition](#)

How were prizes determined?

The LessWrong moderation team looked over the results of the vote, [sliced four different ways](#):

- Total votes from *all users*;

- Total votes from *1000+ karma users*;
- Total votes from *Alignment Forum users*;
- Weighted votes from *all users, but with 1000+ karma users weighted 3x*.

That gave us some sense of how different users related to posts, and what they got out of them. In all cases, approximately the same posts showed up at the top of the rankings, albeit in somewhat different orders. I'm excited to reward all of them prizes for contributing to important intellectual progress.

Two things stuck out between the various rankings:

- In the Alignment Forum vote, [Draft Report on AI Timelines](#) won by a landslide.
- In the "All Users" vote, the [MicroCOVID.org announcement](#) won by a landslide.

Thoughts on the microCOVID.org Announcement

Several people asked, upon seeing microCOVID.org in the review: "Is this what the Review is meant to reward? It sure was useful, but was it 'intellectual progress'? Was it 'timeless'? What would it mean to put it in a book?"

We haven't yet made our final call on whether/how to do books this year, but if I were to print microCOVID, I'd specifically be interested in printing the [whitepaper](#). Unlike the announcement (which would be quite weird to read in a book), the whitepaper is clearly a piece of enduring intellectual labor.

I actually took some pretty timeless, valuable lessons from microCOVID – the schema that you can break "risk" into units that linearly add up was helpful for my overall thinking, and I expect to be relevant in other domains. And the conceptual breakdown of splitting up "person risk" and "activity risk" was useful.

I think microCOVID was a great instance of applied rationality that built off a lot of concepts in our ecosystem, and translated them into a concrete product.

Thoughts on Draft Report on AI Timelines

Ajeya's Draft Report on AI Timelines inspired a ton of discussion: Eliezer [wrote a response post](#), and Holden wrote a [counter-response](#), and there was further discussion on each post. I plan to include both follow-up posts as contextual reviews in the Best of 2020 sequences and/or books.

Daniel Kokotajlo wrote this review articulating why the post seemed important to him (edited slightly for brevity):

Whenever people ask me for my views on timelines, I go through the following mini-flowchart:

1. Have you read Ajeya's report?

– If yes, launch into a conversation about the distribution over 2020's training compute and explain why I think the distribution should be [substantially to the left](#), why I worry it might shift [leftward faster](#) than she projects, and why I think we should use it to forecast [AI-PONR](#) instead of TAI.

– If no, launch into a conversation about Ajeya's framework and why it's the best and why all discussion of AI timelines should begin there.

So, why do I think it's the best? In a nutshell: Ajeya's framework is to AI forecasting what *actual climate models* are to climate change forecasting (by contrast with lower-tier methods such as "Just look at the time series of temperature over time / AI performance over time and extrapolate" and "Make a list of factors that might push the temperature up or down in the future / make AI progress harder or easier," and of course the classic "poll a bunch of people with vaguely related credentials.")

Ajeya's model makes only a few very plausible assumptions. This is underappreciated, I think. People will say e.g. "I think data is the bottleneck, not compute." But Ajeya's model doesn't assume otherwise! If you think data is the bottleneck, then the model is more difficult for you to use and will give more boring outputs, but you can still use it.

I think a lot of people are making a mistake when they treat Ajeya's framework as just another model to foxily aggregate over. "When I think through Ajeya's model, I get X timelines, but then when I extrapolate out GWP trends I get Y timelines, so I'm going to go with $(X+Y)/2$." I think instead everyone's timelines should be derived from variations on Ajeya's model, with extensions to account for things deemed important (like data collection progress) and tweaks upwards or downwards to account for the rest of the stuff not modeled.

The Unit of Caring prize pool

Twelve users donated to the Unit of Caring prize pool. \$1375 was donated to thank specific authors, and \$395 was donated to the "general moderator discretion" fund. Lightcone Infrastructure matched all the donations, and the LessWrong moderation team reviewed and allocated them as follows:

- \$700 to **catherio** for [microCOVID.org, a Tool for Estimating Covid Risk](#)
- \$500 to **Zvi** for [The Road to Mazedom](#), and [Covid-19: My Current Model](#)
- \$420 to **Johnswentworth** for [When Money is Abundant, Knowledge is the Real Wealth](#), and [Alignment By Default](#)
- \$300 to **Duncan Sabien** for the [CFAR Participant handbook](#)
- \$250 to **Daniel Kokotajlo** for [Against GDP as Metric For Timelines](#)
- \$250 to **abramdemski** for [Most Prisoner's Dilemmas are Stag Hunts, Most Stag Hunts are Schelling Problems](#)
- \$250 to **Jeffrey Ladish** for [Nuclear War Unlikely To Cause Human Extinction](#)
- \$140 to **Mark Xu** for [The First Sample Gives the Most Information](#)
- \$140 to **Richard Korzekwa** for [Why Indoor Lighting is Hard To Get Right and How To Fix It](#)
- \$100 to **evhub** for [An Overview of 11 Proposals for Building Safe Advanced AGI](#)

Prizes for Reviewers

The LessWrong Review isn't just about posts. It's also about the process of reflection and evaluation of those posts. We've been awarding prizes for reviewers who contributed significant commentary on posts. The totals came to:

- \$600 to AllAmericanBreakfast
- \$500 to johnswentworth

- \$400 to Vanessa Kosoy
- \$200 to Steven Byrnes
- \$200 to Zvi
- \$100 each to abramdemski, adamzerner, CharlieSteiner, Daniel Kokotajlo, Davidmanheim, Elizabeth, magfrump, MondSemmel, Neel Nanda, niplav, nostalgebraist, philh, Richard Korzekwa, Turntrout, Vika, Yoav Ravid, and Zack_M_Davis

I separately hired Bucky to do some in-depth reviews.

There were a few different ways I found reviews valuable.

First, often they simply reminded me a good post existed, and gave some context for why it mattered. (johnswentworth's comments on [Subspace Optima](#) were a short and sweet version of this. Daniel Kokotajlo's comments on [11 Proposals for Safe Advanced AI](#) were also a great example, which I expect fed into that post ranking highly in the overall vote)

Second, there were many self-reviews by authors who had learned a lot. Of these, some of my favorites are:

- Steve Byrnes on [Inner Alignment in Salt Starved Rats](#), where he exhaustively goes over some updates he made to his model.
- Niplav's self-review of [Range and Forecasting Accuracy](#) (he expressed worry about a number of mistakes in the post, but I think the review is a great time to fix mistakes on otherwise good posts, and I was generally quite impressed with his thoroughness and thought the topic was quite important).

Third, critical reviews.

- I thought TurnTrout's comment on [Nuclear War Is Unlike to Cause Human Extinction](#) was succinct, and helped draw attention to a missing piece of the story. Landfish responded with some clarification.
- Bucky's commentary on [Kelly Bet On Everything](#) helped iron out where the math exactly worked.
- AllAmericanBreakfast dug into the paper Scott cites in [Studies on Slack](#), which taught me some new empirical things about the world as well as feeding into a comprehensive critique.

Fourth, in-depth reviews that explored an idea and thought about its applications.

- [Johnswentworth](#) and [Vanessa](#) both had good reviews of The Solomonoff Prior is Malign, which is a concept I've found confusing to think about.

My sense is that reviewing generally feels less glamorous than writing top level posts, but I think it's quite important for grounding the overall process.

To all our reviewers, thank you!

One More Thing...

That's *almost* it for this year's review. We have at least one additional bit of work, which is assembling the top posts into sequences. The winning posts will be

integrated into our site library, promoted heavily for new users to read and reflect on.

Congratulations to the winning authors – thank you for contributing to LessWrong's longterm intellectual progress.

1. ^

During the Review Phase, people were encouraged to write comments reviewing each nominated post, reflecting on why it was valuable or how it could be improved.

The Big Picture Of Alignment (Talk Part 1)

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

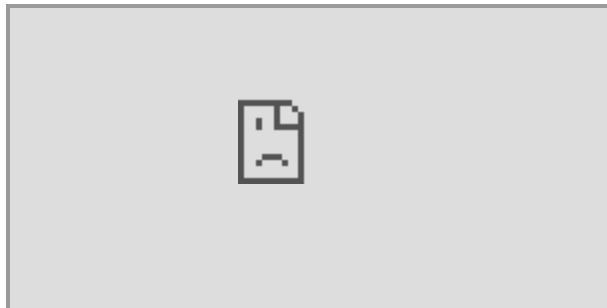
This is a linkpost for <https://www.youtube.com/watch?v=mij7nYPKIHo>

I recently gave a two-part talk on the big picture of alignment, as I see it. The talk is not-at-all polished, but contains a lot of stuff for which I don't currently know of any good writeup. Major pieces in part one:

- Some semitechnical intuition-building for high-dimensional problem-spaces.
 - Optimization compresses information "by default"
 - Resources and "instrumental convergence" without any explicit reference to agents
- A frame for thinking about the alignment problem which only talks about high-dimensional problem-spaces, without reference to AI per se.
 - The central challenge is to get enough bits-of-information about human values to narrow down a search-space to solutions compatible with human values.
 - Details like whether an AI is a singleton, tool AI, multipolar, oracle, etc are mostly irrelevant.
- Fermi estimate: just how complex are human values?
- Coherence arguments, presented the way I think they *should* be done.
 - Also subagents!

Note that I don't talk about timelines or takeoff scenarios; this talk is just about the technical problem of alignment.

Here's the video for part one:



Big thanks to Rob Miles for editing! Also, the video includes some good questions and discussion from Adam Shimi, Alex Flint, and Rob Miles.

Bryan Caplan meets Socrates

Socrates and Glaucon are walking down from the Acropolis, when they encounter a stranger from a distant land.

Caplan: Greetings, Socrates.

Socrates: Greetings, stranger. From whence do you come?

Caplan: I am from a faraway land.

Socrates: Sparta? Thrace?

Caplan: Much further out than that.

Socrates: Where, then?

Caplan: It is not important right now. I have heard that you are the wisest man in Athens, and I have sought your expertise. Socrates, what is the purpose of education?

Socrates: To refine virtue, of course.

Caplan: And so those with an education are more virtuous than those without?

Socrates: Yes.

Caplan: Is it not true, then, that those with an education will be entrusted with greater responsibilities? That they will be made rulers, put in charge of important military expeditions, and will be respected craftsmen?

Socrates: Of course.

Caplan: After a time, will men not seek out an education just for these good consequences?

Socrates: They surely will. It would be better if they sought education for its own sake. Some men will seek it for its good consequences, but at least some will refine their virtues in doing so.

Caplan: What if sophists took over the academies, and no longer taught virtue at all? Men would learn nothing of import, and only become educated to enter the skilled professions.

Socrates: No one would trust such academies.

Caplan: Perhaps. But what if the academies taught both virtue and sophistry? Would the self-interested man not take lessons so as to give the *appearance* of virtue, while exerting himself to the minimum extent? And imagine, Socrates, that you are employing a skilled professional. Would you not employ that man with the greatest education?

Socrates: I surely would.

Caplan: Is it not the case, then, that to the professionals looking for workers, *it does not matter whether they had a valuable education?* It only matters that their education *signals* them to be good workers, who will show up on time and work to their greatest extent?

Socrates: It appears so.

Glaucon: Your words are indeed convincing, traveller. However, I do not see their import. Athens is the most learned city of them all, and even here boys are educated only for a few years. Boys will not sit around learning sophistry if there are wars to fight, or if there is food to grow.

Caplan: That is no doubt true, Glaucon. However, consider this: a ruler will be popular if he supports education. The people are not trained in philosophy, and they cannot follow the argument I have given you. And if they can, they do not wish to.

Socrates: The purpose of a ruler is not to be popul-

Caplan: Yes, yes! But it is only natural for a ruler to desire to be liked by his citizens.

Socrates: The education of a ruler should rid him of such desires, as I discussed before with Glaucon.

Caplan: Have we not already said that the academies can be infiltrated by sophists?

Socrates: I know of no such academy that philosophers respect.

Caplan: But men in the military and the skilled professions are not philosophers. They must rely on crude appearances, to save time. But let us put this aside for the moment. Rulers will be popular if they support education. They will also have been told from a young age that education instils virtue, even if it does not. Teachers themselves stand to gain a great deal from maintaining the prestige and wealth that rulers grant them. Rulers therefore will give much more wealth and esteem to education than it deserves.

Glaucon: Rubbish!

Socrates: Glaucon, restrain yourself! Our traveller has proved himself to be philosophically learned. But it is getting dark, and Glaucon must return home. I will think this over and we will discuss it in the morning.

The next evening.

Caplan: Socrates, I have been looking for you. I have visited the priestess at Delphi, and she has told me of her premonition about education.

Glaucon: Impossible! How have you returned to Athens so quickly?

Caplan: Never mind that for now.

Socrates: What did she say?

Caplan: She said that, in the land from which I come, boys (and girls too!) will eventually be educated for as many as seventeen years. Those entering advanced professions may study for more than twenty. They will not exert themselves in the course of their studies, but instead, drink wine and play games. The academies will be

luxurious, with [man-made rivers](#) flowing through them. They will be treated like royalty, paid for by a tithe on working men. Things will not be much better in Athens.

Socrates: This is one of the most absurd prophecies I have ever been told, but the Delphic Oracle does hold much wisdom... Will your land contain bountiful riches, such that every man lives like a king?

Caplan: Somewhat. However, the gains will particularly go to those that study at the academies. They will gain almost twice as much silver after their studies.

Glaucon: And so, what fool would not study there?!

Caplan: Indeed, most men of wisdom do. But others leave the academy because they find it so *boring*.

Glaucon: Boring?! Socrates, this is a strange traveller indeed...

Caplan: Let me explain! Socrates, the priestess told me that your method of instruction spreads far and wide for a time, but then dies out when the use of writing becomes common. It is replaced with a form of instruction that induces sleepiness and, at worst, contempt for the subject being studied.

Socrates: Plato, I told you so!

Plato [scribbling furiously]: Hey!

Glaucon: Never mind all that. Will the craftsman and workers not realise that this situation is absurd, and rebel? You said yourself that popularity is important to a ruler, even if his education is supposed to get rid of such concerns.

Caplan: Alas, the system is popular even among them! There will be one handsome fellow, a philosopher of sorts, who [points out the absurdity](#), but his ideas will receive little attention among rulers.

Glaucon: How odd.

Socrates: While you were speaking, I was thinking over this prophecy, and I have a few explanations. First, the professions of the future may be more complicated, and therefore require many more years of study. For example, ships will be able to travel farther, but only because shipbuilders spend many more years as apprentices. As a philosopher, I have had to read only the works of Thales and a few others, but philosophers of the future will have to read much more widely. Second, a certain level of material comfort is required to learn. We Athenians need only the basic comforts, but perhaps men from other lands need more. We know that Phoenicians need more silver than us to live without strife.

Caplan: These are excellent points. However, I have been told that the growth in education is mostly *within* the professions and not *between* them. To build even the same ships requires more years of shipbuilding experience.

Socrates: Be that as it may, there must be some quicker way of giving the appearance of skill, without spending many years in the academy. A contest, perhaps.

Caplan: I thought this also. Rulers from my part of the world [restrict](#) how and when you can run such contests, but I do not think this is so important. More important is

that the academy gives the appearance of many skills. Intelligence, but also timeliness, politeness, and ability to deal with men from other parts of the world. A willingness to do tasks asked of you without questioning them. All of these are important to the professionals, and they are not easily displayed in a contest. And regarding material comforts, I agree that some of them are necessary to think well. However, the material comforts of which the priestess spoke far exceed this. And worse yet, most educated men believe that the academies should receive *more* of their wealth, and not less. Especially philosophers!

Socrates [chuckling]: Excuse me, traveller, but you have tickled me, for I misheard and thought you said philosophers believed the sophistry you have spoken of, and wanted more wealth for the academies.

Caplan: You have not misheard! The philosophers love the academies, because they are showered with praise and esteem for their intelligence and hard work. The bulk who dislike the academies often are not skilled in such areas, and so cannot articulate good objections to the philosophers.

Glaucon: Speaking of material comforts, we are leaving now for dinner and wine with the others. Do you wish to join us, and tell them of the premonition?

Caplan: Sounds great!

The next morning.

Caplan: Good morning, Socrates. I have one more topic on which I seek your counsel. It is true, is it not, that most men have no interest in philosophy, and in such fine arts as poetry?

Socrates: Unfortunately so.

Caplan: And therefore education, insofar as it is given to everyone, should not include these elements?

Socrates: This does not follow. The lack of interest in philosophy and the fine arts only shows that people have not received sufficient instruction to awaken their love for it.

Caplan: And what makes you so confident that we all have such a love, waiting to be awakened?

Socrates: As I explained last night at dinner, it is because of the tripartite nature of the soul. Our soul separately houses intellectual, emotive and appetitive pleasures. This is the only way we can account for the paradox of opposites. Love of wisdom, therefore, is part of the soul.

Caplan: People from my country have very different views on this subject, but let us put that aside. Do you think this love of wisdom can be awakened in all people, even women and slaves?

Socrates: Huh, I had not previously considered women and slaves...

Caplan: While you think, I shall tell you more about the premonition I was told at Delphi. In the future, every girl and boy will be instructed in fine arts and disciplines like philosophy, literature, and poetry. Whether or not their interest *can* be awakened,

it is not in almost all cases. Teachers with love for their subjects flee into other professions, and this leads to a chicken-and-egg problem. If the students are uninspired because of bad teachers, and good teachers will not teach uninspired students, how do you fix that?

Socrates: Chicken-and-egg... That's a humorous comparison... I may use that.

Enter Thrasymachus.

Socrates: Thrasymachus, our friend here is talking about awakening the love of knowledge in students. If students are uninspired, then only uninspiring teachers will choose to teach them.

Thrasymachus: This is perhaps true. But consider this: students may show promise in other ways. The skills gained in philosophy and poetry sharpen the mind, and teach you how to think, even if you do not love them for their own sake. These skills may be applied to other areas. And educators teach those who show promise in any area. For example, I mentored a boy as a favour to a friend. I was reluctant at first, but the boy was a prodigious mason. I saw promise and applied myself to him. At the end of our time, I saw in him the beginnings of a love of philosophy and the arts.

Caplan: My contention is only that such cases are rare. Socrates, can a youth not go to the Acropolis and hear all manner of ideas about philosophy?

Socrates: Yes, he can.

Caplan: And yet youths do not go, as a rule. Why is that?

Socrates: Because they have no interest.

Caplan: And consider also this. Thrasymachus, does training as a stonemason make you a better shipbuilder?

Thrasymachus: Surely not, except in the broadest ways of using some tools.

Caplan: Precisely. The transfer is there, but it is limited. So: why does learning poetry make you a better stonemason? Shipbuilding is surely more similar to masonry than to poetry, is it not?

Thrasymachus: Poetry and philosophy refine the *mind*, and the mind can be applied to anything. While shipbuilding only refines the *hands*, and the body, which can only be applied to certain tasks.

Caplan: Excellent, Thrasymachus. I just have one question: what makes you confident that the mind is a single entity, where training one part of it trains the entire thing? If you train your hands through pottery, that does not train your legs for running, merely because they are both parts of the body. Perhaps poetry only refines the poetry part of the mind.

Thrasymachus: The mind is unified because we can exert a will. When you exert yourself toward a goal, you will use every skill that your mind is capable of. But the body cannot exert a will. When the body moves in a coordinated fashion, it is only because our mind is controlling it. An unconscious man cannot move in a coordinated way.

Caplan [aside]: Wow, Athens really doesn't have sleepwalkers?

Caplan: Very well, Thrasymachus. This issue is complex, and I must return home soon. Socrates, I have one last thing to ask of you. I worry that knowledge from philosophy and the arts is only learned in theory, and not in practice, thus not justifying the large public expense of which the Oracle spoke. For example, when visiting the temple, philosophers [do not pay a tribute](#) at any greater rate than men of similar social standing. I love the realm of ideas, Socrates, and this is why I have travelled so far to speak with you. However, most men don't. And I fear that learned rulers enforce their interests on the rest of the populace, and that this is an incalculable waste of time and wealth.

Socrates: If what you speak of is true, I admit it is troubling. Perhaps philosophy is what *allows* men to live ethically, but on average does not change their behaviour. I always pay a tribute upon visiting the temple.

Thrasymachus: You already know my views on justice, but it is commonly said that Socrates is the most just man in all of Athens.

Caplan begins packing up his bags to leave.

Socrates: You have certainly given us much to think about, traveller. And I see now that you must return home. I don't wish for you to carry those heavy bags by yourself, so I will send a slave with you.

Socrates calls out for a slave.

Caplan: No, it is fine! Thank you, Socrates, this has been a most informative visit. Send my best to Xanthippe.

Caplan leaves.

Thrasymachus: What a strange fellow.

Socrates: Indeed, Thrasymachus, indeed.

The Territory

Knowing the territory takes patient and direct observation.

I don't know if you've ever had the experience of hiking in the wilderness with a map and compass but no cell service. I recommend it, if you haven't.

I somehow did not quite all-the-way understand what a map even is until I was lost on my own under these circumstances. I knew in a "factual knowledge" sense that maps were drawings of the land, and I'd even used them as a kid and teenager to help my family navigate on road trips. But when I was lost in a national park, trying to find my way back to my car, I confronted the incompleteness of my knowledge of maps. There was a shift.

My map had trail lines drawn on it, with labels like "Canyon Trail". I'd pause my walking to look at the shape of "Canyon Trail", noting that it intersected "Overlook Trail" somewhere off to the left of where I was standing. Then I would walk again—attempting, I think, to "follow Canyon Trail to Overlook Trail".

I would move back and forth between walking and map consultation, making sure I remembered which way the trails were supposed to go, constantly placing and replacing myself within the borders of the lines drawn on the paper. The more distressed I felt about being lost, the more often I turned to the map, looking for something to hold on to.

The shift happened after... (this is sort of embarrassing, it's so simple. But it's true.) The shift happened after, having oriented myself toward "North", I happened to lower the map a little bit, probably out of exhaustion. I held it a bit below eye level, so that it was no longer taking up my whole field of vision.

I looked at the squiggly blue line on the map, and the close-together lines that I knew indicated steepness. And I saw to my left, because the map was not blocking my vision, a creek. Up ahead, I saw a steep hill.

I realized that the blue line was probably a drawing of *that creek*.

The contour lines were a drawing of *that hill*.

And then this wild rushing sensation began to wash over me. I was starting to get it. Slowly, I tilted the paper in my hands from a vertical position, partially blocking my view...

...to a horizontal position, parallel to the ground.

I held the map that way, looking out at the world the cartographer had tried to draw, and it was as though the territory rose up to meet the map, while the map spread itself across the surface of the territory. And I said to myself, "It's a picture!"

For the first time, I understood in a practical way that a map is meant to be a top-down picture of the real world.

Before I had this realization, I wasn't behaving as though I knew myself to be in the territory, using the map as a tool. I was acting as if I were traversing the map, using my body as a kind of clunky video game controller. I had been treating *the map* as the terrain I "really" had to navigate.

But once I stopped playing that game, and started actually traversing the forest I was in, things went very differently. I spent most of my time looking at creeks and trees and hills, making sure I knew how the real world around me was shaped. And from *that* perspective, I looked down at the map to help me predict what I'd see next.

And I found my car shortly thereafter.

There are ways to increase some kinds of knowledge that largely involve staring at maps. Perhaps your own map is not clearly labeled in places, or it's somehow inconsistent with itself, or it doesn't match the map of an expert.

This is why it's often valuable to clearly articulate your beliefs, even just to yourself. It's valuable to ask yourself what you expect, and to notice when you feel confused about that. It's valuable to ask other people what they think, or to read their books and blog posts, especially when you have reason to believe they know important things that you don't.

But the *main* thing a cartographer ought to be focused on, the vast majority of the time, is the world itself.

I started studying "[original seeing](#)", on purpose and by that name, in 2018. What stood out to me about my earliest exploratory experiments in original seeing is how *alien* the world is.

I don't mean that reality is weird or surprising. Nothing weird has ever happened, and all of that. What I mean is... well, I think I should actually grab [an Eliezer quote](#) here:

Human intuitions were produced by evolution and evolution is a hack. The same optimization process that built your retina backward and then routed the optic cable through your field of vision, also designed your visual system to process persistent objects bouncing around in 3 spatial dimensions because that's what it took to chase down tigers. But "tigers" are leaky surface generalizations - tigers came into existence gradually over evolutionary time, and they are not all absolutely similar to each other. When you go down to the fundamental level, the level on which the laws are stable, global, and exception-free, there aren't any tigers. In fact there aren't any persistent objects bouncing around in 3 spatial dimensions.

I started my earliest experimentation with some brute-force phenomenology. I picked up an object, set it on the table in front of me, and progressively stripped away layers of perception as I observed it. It was one of these things:



I wrote, "It's a SIM card ejection tool."

I wrote some things about its shape and color and so forth (it was round and metal, with a pointy bit on one end); and while I noted those perceptions, I tried to name some of the interpretations my mind seemed to be engaging in as I went.

As I identified the interpretations, I deliberately loosened my grip on them: “I notice that what I perceive as ‘shadows’ needn’t be places where the object blocks rays of light; the ‘object’ could be two-dimensional, drawn on a surface with the appropriate areas shaded around it.”

I noticed that I kept thinking in terms of what the object is *for*, so I loosened my grip on the utility of the object, mainly by naming many other possible uses. I imagined inserting the pointy part into soil to sow tiny snapdragon seeds, etching my name on a rock, and poking an air hole in the top of a plastic container so the liquid contents will pour out more smoothly. I’ve actually ended up keeping this SIM card tool on a keychain, not so I can eject SIM trays from phones, but because it’s a great stim; I can tap it like the tip of a pencil, but without leaving dots of graphite on my finger.

I loosened my grip on several preconceptions about how the object behaves, mainly by making and testing concrete predictions, some of which turned out to be wrong. For example, I expected it to taste sharp and “metallic”, but in fact I described the flavor of the surface as “calm, cool, perhaps lightly florid”.

By the time I’d had my fill of this proto-exercise, my relationship to the object had changed substantially. I wrote:

My perceptions that seem related to the object feel very distinct from whatever is out there impinging on my senses. ... I was going to simply look at a SIM card tool, and now I want to wrap my soul around this little region of reality, a region that it feels disrespectful to call a ‘SIM card tool’. Why does it feel disrespectful? Because ‘SIM card tool’ is how I use it, and my mind is trained on the distance between how I relate to my perceptions of it, and what it is.

There aren’t any tigers, and there aren’t any SIM card tools, either. It now feels... almost *disgusting*, to me, to lose sight of that. Disgusting like thinking of trees only as “lumber”, and cutting down entire rainforests as a result.

Which doesn’t mean it’s useless to conceptualize tigers and so forth. It absolutely is useful and correct. The purpose of cartography is to draw cartoon pictures that are relatively useful to travelers, and certain features of the cartoon pictures need to correspond to the real-world not-actually-“tigers” to be useful. There exist for-real regions (or properties, or patterns) of the territory itself that it *makes sense* to call “tigers”, as long as that concept is doing the right stuff, such as paying rent in anticipated experiences.

But ever since I began my study of original seeing—ever since observing the so-called “SIM card tool”—it has felt a little different for me to use the word “territory”.

I think that before, when I said “the territory”, I must have accidentally meant something like “the much bigger map; the thing I’m drawing a map of, which is basically like my map but a lot more complex”.

Now I mean something like, “The thing that is made of something other than my own perceptions and interpretations. The thing that resists my expectations, according to its own rules. The thing that does not care what I think, or what I have happened to imagine.”

In the sentence, “Knowing the territory takes patient and direct observation,” what I mean by “territory” is “the thing that is made of something other than my own perceptions and interpretations”.

Knowing [the thing that is made of something other than your own perceptions and interpretations] takes patient and direct observation.

Next, there will be a short interlude on realness, and what it feels like to lower the map. Then I'll talk about *observation* of the territory.

Compute Trends Across Three eras of Machine Learning

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

<https://arxiv.org/abs/2202.05924>

What do you need to develop advanced Machine Learning systems? Leading companies don't know. But they are very interested in figuring it out. They dream of replacing all these pesky workers with reliable machines who take no leave and have no morale issues.

So when they heard that [throwing processing power at the problem might get you far along the way](#), they did not sit idly on their GPUs. But, how fast is their demand for compute growing? And is the progress regular?

Enter us. We have [obsessively analyzed](#) trends in the amount of compute spent training milestone Machine Learning models.

Our analysis shows that:

- **Before the Deep Learning era**, training compute approximately followed Moore's law, doubling every ≈ 20 months.
- The **Deep Learning era** starts somewhere between 2010 and 2012. After that, doubling time speeds up to $\approx 5\text{-}6$ months.
- Arguably, between 2015 and 2016 a separate **trend of large-scale models** emerged, with massive training runs sponsored by large corporations. During this trend, the amount of training compute is 2 to 3 orders of magnitude (OOMs) bigger than systems following the Deep Learning era trend. However, the growth of compute in large-scale models seems slower, with a doubling time of ≈ 10 months.

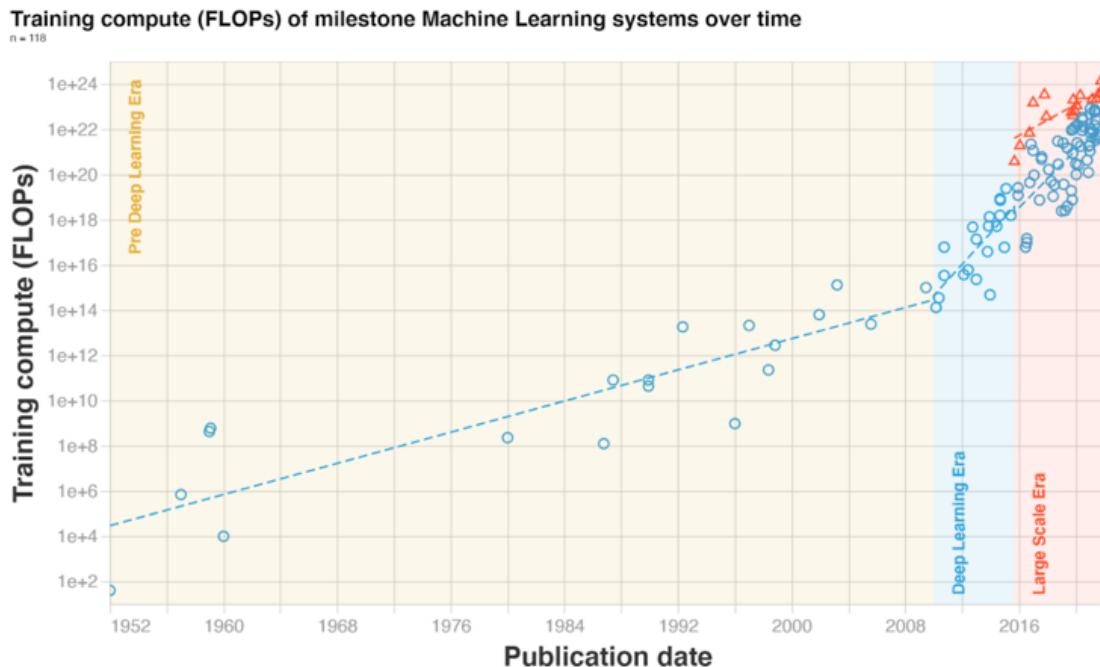


Figure 1: Trends in n=118 milestone Machine Learning systems between 1950 and 2022. We distinguish three eras. Note the change of slope circa 2010,

matching the advent of Deep Learning; and the emergence of a new large scale trend in late 2015.

| Period | Data | Scale (start to end) | Slope | Doubling time |
|------------------------|----------------------------------|----------------------|----------------------------------|-----------------------------------|
| 1952 to 2010 | All models (n = 19) | 3e+04 to 2e+14 FLOPs | 0.2 OOMs/year [0.1; 0.2; 0.2] | 21.3 months [17.0; 21.2; 29.3] |
| 2010 to 2022 | Regular-scale models (n = 72) | 7e+14 to 2e+18 FLOPs | 0.6 OOMs/year [0.4; 0.7; 0.9] | 5.7 months [4.3; 5.6; 9.0] |
| September 2015 to 2022 | Large scale models (n = 16) | 4e+21 to 8e+23 FLOPs | 0.4 OOMs/year [0.2; 0.4; 0.5] | 9.9 months [7.7; 10.1; 17.1] |
| Large-Scale Trend | | | | |

Table 2: Summary of our main results. In 2010 the trend accelerated along with the popularity of Deep Learning, and in late 2015 a new trend of large-scale models emerged.

Table 1. Doubling time of training compute across three eras of Machine Learning. The notation [low, median, high] denotes the quantiles 0.025, 0.5 and 0.975 of a confidence interval.

Not enough for you? Here are some fresh takeaways:

- Trends in compute are **slower than previously reported!** But they are **still ongoing**. I'd say slow and steady, but the rate of growth is blazingly fast, still doubling every 6 months. This probably means that you should double the timelines for all **previous analyses** that relied on *AI and Compute's* previous result.
- We think the framing of the **three eras of ML** is very helpful! Remember, we are suggesting to split the history of ML into the **Pre-Deep Learning Era**, the **Deep Learning Era** and the **Large-Scale Era**. And we think this framing can help you make sense of what has happened in the last two decades of ML research.
- We have curated an awesome **public database of milestone ML models!** Please use it for your own analyses (don't forget to cite us!). If you want to play around with the data, we are maintaining an interactive visualization of it [here](#).

Compute is a strategic resource for developing advanced ML models. Better understanding the progress of our compute capabilities will help us better navigate the advent of transformative AI.

In the future, we will also be looking at the other key resource for training machine learning models: *data*. [Stay tuned for more!](#)

[**Read the full paper now on the arXiv**](#)

Observations about writing and commenting on the internet

This is a linkpost for <https://dynamight.net/internet-writing/>

I'm not famous or successful, so why should you care what I think? Well, I have some observations about the dynamics of writing on the internet that I think my (even more non-famous and non-successful) self would have benefited from when I started.

Human experience is vast.

The whole idea of writing is crazy: You have a pattern in your brain-meat, which you try to encode it into a linear series of words. Then someone else reads those words and tries to reconstruct the pattern in their brain-meat. But in this dance, how much of the work is being done by the words versus the lifetime of associations each person has built up around them?

Rather than a full blueprint for an idea, writing is often more like saying "Hey, look at concept #23827! Now look at concept #821! Now look at concept #112234! Are your neurons tingling in the way mine are? I hope so because there's no way to check, bye!"

We have different personalities and spend our lives getting exposed to different information and thinking about different things. What concept #821 triggers for you may be vastly different than what it triggers for me.

I suspect that even when writing works, readers are often taking quite a "different trip" than the writer intended. Personally, I figure that's fine and it's better to just *let* people take their own trip, rather than going to insane lengths in a hopeless quest to make everything precise.

So: No matter what you do, sometimes your writing will fail. It's impossible to predict all the ways it will fail. Really, it's amazing that it works at all. But still, it's possible to reduce the frequency of failure.

Failures can seem baffling.

Here are two examples of how things I've written have failed:

1. I wrote an article suggesting ultrasonic humidifiers might put particulates into the air and harm health. The median response was something like this:

"Please stop polluting the internet with speculation. If you can't support your argument with peer-reviewed research papers you shouldn't write it at all."

This was... puzzling, because I had dozens of citations starting very early in the article. But many people in different forums *independently* left comments like this.

2. I wrote an article about gas stoves and nitrogen dioxide where I stated in the overview that a normal range hood might not solve the problem. A top comment was that a *high-quality* hood would solve the problem and the fact that I claimed this without providing any data showed I was biased and bad (so bad).

That's fair, although after the overview finished there was a link to the "Can a range hood fix this?" section where I discuss an article that tested hoods in various houses and found that some worked well but most of them reduced nitrogen dioxide by 30% or less, and wait did I say that was fair?

Please holster your [tiny violin](#), I know this is all normal. Many people get fed up with stuff like this and resolve to ignore all comments. That's not my position. My position is that I failed and when I fail I'd like to know about it.

First, though, why does this happen?

What happens in forums isn't always about you.

Why do people go to forums? Partly for links, yes, but also because they like the community. When an article on an interesting topic comes up, some people decide they'd rather see what people they trust think before investing time reading the article. And once they are there, maybe they see a comment they want to respond to.

I'd love to see statistics for this, but I'd guess that only a small fraction of people finish most articles before commenting, and many don't even look at the article. This is not necessarily a bad thing! Sometimes the discussion detaches and goes into all sorts of interesting and unexpected directions.

Sometimes this can be funny, too. Once I wrote about a study by [Pierson et al. \(2020\)](#) who investigated how the racial mix of drivers stopped by police changes at different times of the year when it may be harder to see the drivers. The first comment was basically, "This is all wrong. There's a study by Pierson et al. (2020) that investigated..."

Engagement has a sample bias.

The first few times I saw my posts discussed, I was shaken by the amount of negativity. This bothered me until I tried going to posts from others that I thought were great and reading the comments as if I were the author. Sometimes, umm, the comments were all glowing. But other times—with no clear pattern—most dismissed the article as pointless/obvious/wrong/bad. After this, negativity still bothered me, but I had a new variant of the old "Einstein was bad at math, I'm bad at math" fallacy to distract me.

Here's something I've noticed about myself: If I read something great, I'll sometimes write a short comment like "This was amazing, you're the best!" Then I'll stare at it for 10 seconds and decide that posting it would be lame and humiliating, so I delete it and go about my day. But on the rare occasions that I read something that triggers me, I get a strong feeling that I have *important insights*. Assuming that I'm not uniquely broken in this way, it explains a lot.

Listening to criticism is a superpower.

Do you have a friend who works in user-facing software? Sometime after they've had a few drinks, ask about the first time they saw one of their creations being tested on real people. Observe how somber their face becomes as they express their feelings of frustration and impotence. The things we build are no match by the might of human ingenuity to do everything wrong in unexpected ways.

It's puzzling that there isn't a stronger tradition of "user testing" for writing. Occasionally I'll give a friend something I've written and implore them, "Please circle anything that makes you feel even slightly unhappy for any reason whatsoever." Then I'll ask them what they were thinking at each point. There are always "bugs" everywhere: Belaboring of obvious points, ambiguous phrases, unnecessary antagonistic language, tangential arguments about controversial things that don't matter, etc.

Fixing these is great but your friends (let's hope) don't want to hurt your feelings. This makes it almost impossible to get them to say things like, "your jokes aren't funny" or "you should delete section 3 because it's horrendous and unsalvageable". Good editors are gold.

Comments from the internet are the opposite. A downside is that you have much longer feedback loops, which makes it hard to figure out cause and effect. But you get feedback at a much higher scale and people are... substantially less worried about offending you.

Take the humidifiers example from before. Technically, the complaints were wrong. How could I "fix" the problem of not citing any papers when I had already cited dozens? That's what I thought for months, during which people continued to read the post and have the same damned reaction. Eventually, I had to confront that even if they were "wrong", something about my post was *causing* them to be wrong. Viewed that way, the problem was obvious: The idea that a humidifier could be bad for you is weird and disturbing, and weird and disturbing things are usually wrong so people are skeptical and tend to find ways to dismiss them.

Should they do that?

[Insert long boring polemic on Bayesian rationality]

It's debatable—but it's a fact that they do it. So I rewrote the post to be "gentle". Previously my approach was to sort of tackle the reader and scream "HUMIDIFIERS → PARTICLES! [citation] [citation] [citation] [citation]" and "PARTICLES → DEATH! [citation] [citation] [citation]". I changed it to start by conceding that ultrasonic humidifiers don't *always* make particles and it's not *certain* those particular particles cause harm, et cetera, but PEER-REVIEWED RESEARCH PAPERS says these things are possible, so it's worth thinking about.

After making those changes, no one had the same reaction anymore.

Part of me feels like this is wrong, that it's disingenuous to tune writing to make people have the reaction you want them to have. After all, I could be wrong, in which case it's better if my wrongness is more obvious.

Maybe there's a slippery slope here, but I think most people operate very close to the top of that hill. The goal of writing is to communicate, and it's silly to ignore the effects it has on the actual people who read it.

So my advice is this: When you hear criticism, you need to guess if people even looked at the post. If they did, *some* negative reactions are inevitable. But if you repeatedly hear the same complaint, you should have a strong presumption that there is a problem, though it might be very different from the problem people state.

No one is better than the combined efforts of a large group of people.

If comments are often bad, does that mean you shouldn't read them? If you're very fragile, maybe. But you'll be missing out. For one thing, you can often try to trace back the causal chain like above.

A bigger reason is just that sometimes comments are insanely great. Here's a [comment](#) from blockcipher on a post about methamphetamines:

There's a common myth among tweakers about "n-iso", which is structurally very similar to methamphetamine - similar enough that it will join the crystal lattice - but it is at best inert, but might actually cause undesirable side effects. The fact that n-iso exists is real, but if you look online you'll see tons of tweakers convinced that they've been smoking n-iso and that it's why they smoke meth and just get a headache and other bad physical side effects but don't get the stimulation or the pleasurable rush. What's actually happening is that they've spiked their tolerance so high that they're getting almost exclusively the bad effects. It's analogous to how if someone takes MDMA for 4 days straight, by the end of it they're not going to "roll" at all because they've acutely downregulated their serotonin (and dopamine) receptors, and furthermore that they've literally (almost) exhausted their current pool of neurotransmitters, which need to be re-synthesized by the body.

Or here's a [comment](#) from svat on a post about the proper usage of analogies:

In Indian/Sanskrit literary theory (poetics), in the discussion of figures of speech (rhetoric, etc), similes are called *upamā* ("her face is like the moon", etc). The discussion of it in the literature is extensive and would fill several volumes (and I hardly know anything), but one thing recognized early is that in a simile/analogy, there needs to be a *sādharaṇa-dharma*, a shared property: the point is that there's something in common ("her face is *beautiful*, like the moon") while of course there is going to be a lot that is not (the intended meaning is *not* "like the moon, her face is pockmarked, full of craters", etc). In any given instance, this intended shared property may either be stated explicitly, in which case the simile is called "complete", or left implicit, in which case it's called "partial". Both can be highly effective.

Or here's a [comment](#) from Nameless1995 on a post about if selfhood is real:

I think we often tend to conflate our lived experience of unity with the notion that the whole body has some centralized unitary consciousness. The lived experience is a momentary duration, and it doesn't appear to me as a centralized and exclusive instance of consciousness — there could be multiple others (in the same

body) that are inaccessible to “this” consciousness. Considered as such, mental disorders, DIDs, and split brains are not violations of unity of an instance of consciousness, but would be a result of “de-harmonization” of different instances of consciousness (due to information blockage and other reasons).

The idea of sitting down and finding the One Eternal Truth about anything is a fantasy. The universe has fractal levels of detail in every direction. There are a *lot* of ridiculously smart and well-informed people out there, and some of them will have deeper knowledge and insight about basically every facet of every thought you ever have. If you can motivate the collective hivemind to pay attention to something you care about, you’d be crazy not to listen.

Oddly, it seems to me that discussions that fully detach from the original article are on average *better*. If that doesn’t happen, it’s often because people got stuck arguing about minutia. This also happens for detached discussions of course, but they seem to have a better chance of reaching interesting places.

Aside: Techno-optimism is unfashionable at the moment, but I suspect we still haven’t come close to realizing the potential of even the internet technology of the 1990s. When thousands of people converge on a topic, the collective knowledge *far* exceeds any one person, but our current interaction models don’t do a great job of synthesizing it. It’s a difficult problem, but it’s hard to imagine that in a hundred years we won’t have more effective ways to interact.

The people who will pleasantly engage with you clearly signal their intention to be pleasant.

Sometimes I read a comment and I get a weird feeling, but then I convince myself, “They weren’t *rude*. They are making a sincere comment, and people shouldn’t have to humble themselves and stoke my fragile ego.” So I’ll try to respond.

As far as I can recall, this has never worked. Once after I wrote about the Monty Hall problem, someone curtly stated that I clearly didn’t understand it because I didn’t mention that Monty must choose a non-car door to open *randomly*. I thought about this and replied with an argument that if, say, Monty always chose the leftmost non-car door, everything was still fine. They responded that clearly I didn’t even try to read their comment, I’m just like everyone else who doesn’t get it, plus a lot of math that I found incoherent. I wondered—Am I stupid? Was I missing something? So I wrote 25 lines of Python code to simulate it and verified that this didn’t change the probabilities at all. After I posted that code, my correspondent ~~thanked me for the correction~~ changed nothing, acknowledged nothing, and stopped responding.

There have been many instances where someone wrote to me to say I was wrong, we had a productive back and forth, and they convinced me I was indeed partly or entirely wrong. But in every case, their first message looked like this:

Hello friend, I enormously enjoyed your recent fevered rant on [topic]. However, if I may be so bold I wish to point out errors in paragraphs 1, 2, 3, 5, 12, 17, 20, and 21. [errors] Sadly, these issues render your conclusion not just wrong but incoherent and arguably illegal. Still, you’ve done a great service by writing it and

creating a stimulating discussion. Generations to come will admire you! Yours sincerely, Internet Person.

I exaggerate, but it was always *overwhelmingly obvious* from first contact that they were going to be nice. And people who seemed nice always *were* nice.

The tricky situation is cases where someone is mildly (or un-mildly) rude but *also* makes an intriguing point. After many failures, my policy is now to take their comments into account as much as I can and *maybe* reply with “thanks for your input”, but not to engage or ask follow-up questions.

I’m not sure why things are like this or if this pattern generalizes to other people. But I think everyone needs to build some pattern recognition for this and figure out a policy for when they want to engage.

Distribution, Pareto optimality, and quantity vs. quality.

There’s an argument that most writing has no value. It goes like this: Every hour, more text is produced than you could read in a lifetime. If you can write the *best* piece on a given topic, great, but otherwise we don’t need *more content*. And don’t kid yourself—to write the *best* piece, you’d need to pick a single topic, become a world expert, and spend months polishing the writing. Most writing is just people yelling over each other for their own reasons.

The standard response is to gesture towards [Pareto optimality](#): There’s no “best” article on a given topic because there are many dimensions of quality, which people prioritize in their own ways. Unless another article is better than yours *in every dimension simultaneously* you have the potential to be the best article for *someone*.

That’s a nice thought. But surely it’s significant that we have no *mechanism* for that person to actually find the article that’s optimal for them? (Or maybe Google is really onto something and when you think you want a recipe what you *really* need is pages of SEO-optimized autogenerated gibberish.) To contribute value *in practice*, an article needs to be better than everything else for a decently large slice of the population.

That counter-counter-argument seems strong. Yet, I follow a lot of people who write about lots of different topics and it *feels* like I get value from this. Am I delusional?

I don’t think so, but even if I get value, there could be something else that provides *more* value. Still, I can’t shake the feeling that the people I follow truly are brightening my life on net. I have several hypotheses for why:

First, there are a *lot* of topics, and it’s not *that* hard to be the best. Often this is achieved by virtue of being the *only* article on a topic.

Second, it’s easier to understand writing by people you’re familiar with. They can *get to it* without wasting time establishing context.

Third, people have qualities that are fairly consistent across the stuff they write. If I’m familiar with the concepts someone uses and I get their sense of humor and like the way they choose examples, then lots of the stuff they write can immediately become the best article on a topic *for me*.

Fourth, the distribution problem works both ways. Take a model of the internet as millions of people screaming into the night, with readers just bumping into them at random. In this model, you only need to be above “average” to contribute value. Similarly, because distribution is so poor, writers help with “unknown unknowns”. I had no idea I wanted to learn about [Ryszard Kapuscinski](#) before Matt Lakeman wrote about him.

So here’s a thought experiment: What would things be like if you could plug your brain into a robot and automatically get whatever content is closest to your needs? On the margin, there would be less need to “follow” people, and more opportunity for “weirdness”. But it’s unclear what effect this would have on the reach of domain experts versus generalists. I think that comes down to how much we value information versus other qualities like shared context, readability, familiarity, and aesthetics.

Knowing

I'm going to draw some practical distinctions among types of knowledge, as an attempt to tap into your intuitions and avoid having to give some convoluted, ivory-tower definition of the word. I request that you try not to get distracted by where I've drawn my distinctions—the precise placement of the borders is not the point of the exercise. The point of the exercise is to shine your attention on the richness, depth, and complexity of your capacity to know—that the word "know" means more than one thing.

Let's begin with a little formative assessment.

- **Do you know what comes out of your kitchen faucet?**
- **Do you know what glaciers are made of?**
- **Do you know what those big white fluffy things in the sky are?**

If so, then you are *familiar* with water. When someone talks about it, you're not completely lost.

- **Do you know at what temperature water boils?**
- **Do you know the atomic composition of an ordinary water molecule?**
- **Do you know what percentage of your body's volume is water?**

If so, then you know some *facts* about water. Your concept of water contains (at minimum) a few isolated pieces of accurate information.

- **Do you know the way water sounds when it pours into a cup?**
- **Do you know how water feels when it runs over your skin?**
- **Do you know the look of a stream's surface as it glitters in the sun?**

If so, then you are able to *identify* water when you encounter it in real life. You have direct, experiential data. You are able to predict how various encounters with water will impact your senses, and you probably recognize water when those sensations occur (at least sometimes).

- **Do you know what happens if you leave a beer bottle in the freezer overnight?**
- **Do you know how a water mill grinds grain into flour?**
- **Do you know why it rains?**

If so, then you probably have at least one *model*/^[1] of water.

Whether or not that model is explicit, it includes enough structure that you can predict the behavior of water in various situations, even if those situations are outside of your own direct experience. Your model might be rudimentary, in which case the above questions probably produced hesitation or "sort of?" and you'd maybe only be able to produce a short

paragraph in response to each. Or your model might be rich and deep, in which case your “yes” was confident, and you could in principle write multiple essays on the subject.

- **Do you know how to swim?**
- **Do you know what to expect when applying watercolor paints to a wet canvas?**
- **Do you know how to make sea water safe to drink?**

If so, then you have some *practical mastery* of water. It’s not just that you recognize water, or that you know some things about it, or that you can predict its behavior—your models of water are *integrated* with your models of yourself and other parts of the world, accurately and deeply enough that when you personally interact with water in real life, things tend to go more or less as you intend (at least in certain kinds of situations).



Breaking the format of the pop quiz now, to ask a more difficult question:

Can you name other things you know as intimately and thoroughly as water? Is there some swath of the territory with which you have extensive familiarity, lots of factual knowledge, rich predictive and explanatory models, *and also* practical mastery in a wide variety of situations? In other words, where do you think you might have *deep mastery*?

One such domain that’s likely common: many people have this portfolio of knowledge when it comes to driving cars. The average American spends [a little under an hour a day in a car](#), so if you’re like the average American in this respect (and also you’re the one doing most of the driving for your household), then you’ve plausibly spent over three thousand hours behind the wheel in the past ten years. And if so, I expect you’ve deeply mastered driving^[2].

If you’ve deeply mastered driving, then you have extensive familiarity with all sorts of driving-related tasks and phenomena. You recognize left-turn-only lanes, brake pedals, stop

signs, curves in the road, the hazard lights button on your dashboard, erratic driving, potholes, high beams, deer, and so on.

You probably have tons of factual knowledge related to driving, as well, even if it's been many years since you've taken a written test. If an inquisitive fourteen-year-old were sitting in the passenger seat, you could produce all sorts of relevant bits of data, such as what speeds they should expect to drive on what kinds of roads, or what fluids they'll need to put into their car at what frequencies, or how many wheels most cars have, or what papers they'll need if they get pulled over or have a minor accident.

You probably also have rich, complex models of driving itself, organized to allow you to make reasonable predictions about driving-related situations and phenomena. If your car breaks down on the road, you might or might not know how to fix it, but I bet you at least pull over to the side, because you know how roads work, and you know implicitly that if you stay put, other cars will come up behind you at high speeds and possibly crash into you. If I offered you a large amount of money to fill a hundred pages on "how driving works," you could almost certainly do it, especially if I provided helpful prompts like "differences between driving in cities vs. driving in rural areas" or "things that other drivers frequently get wrong."

And all of these different kinds of knowledge—facts, familiarity, implicit and practical models—they're all *seamlessly integrated* with an experienced driver's knowledge of themselves, and with their knowledge of adjacent domains like travel, geography, weather, car maintenance, the side effects of medication, etc. An experienced driver doesn't (usually) access their knowledge about driving via explicit lookup. They *can* do that, on request, but most of the time they simply *move through the world*. They use their turn signal reflexively in the middle of deep conversation with their passengers. When someone suddenly swerves into their lane, they decelerate without *deciding* to decelerate. They stop for gas on road trips. They notice when something about their car just *feels off*. And they acquired most of this knowledge *in the process* of developing the skills and habits required to safely operate the vehicle.

This is what I mean by "knowing," in the sentence "Knowing the territory takes patient and direct observation." By "knowing", I mean something like *deep mastery*.

[If you want extensive familiarity, accurate factual knowledge, richly detailed predictive models, *and* thorough practical mastery of some part of] the territory (that is, if you want deep mastery), then you will have to engage in patient and direct observation.



1. ^

There is a principled distinction to be made between models and theories. I'm not making it here.

2. ^

If you're doing just fine and enjoying this essay so far, **skip this footnote**. Otherwise, I have some bonus words that might possibly help.

It was around this point when some of my beta readers noticed their frustration with how slowly we were going. They found themselves falling out of the spacious, expansive mode I had hoped they would be in. I think this is fine to do, as long as you think falling out of that mode is a good idea.

But I note that there is another thing you could do, if you're frustrated or bored, which is to look at your own mind, notice the reactions that are happening, ask yourself what they're happening in response to, and thereby ease back into wondering.

Many of us have words for the kinds of distinctions I'm trying to draw here, such as "S1 vs S2", or "tacit vs explicit knowledge", or "declarative vs procedural knowledge". And the thing about those distinctions is that they are a) useful, and b) curiosity-stoppers. They tell us "don't worry, you already know this" so you can get back to building a tower of interconnected concepts. Which is a good thing, most of the time, but it is a bad thing some of the time, and I expect that many of my readers do not know how to tell the difference. (I often do not know how to tell the difference.)

That is (in part) why we are going slowly here, and feeling our way forward without much reliance on a large preexisting vocabulary. That large preexisting vocabulary is good, but it is not perfect. In order to see its flaws, you have to be able to stand outside of it somehow. I'm trying to help you step at least a little bit outside of it.

Why I'm co-founding Aligned AI

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

I'm leaving the Future of Humanity Institute, the best and most impactful job I've ever had, to co-found [Aligned AI](#). For the first time in my research career, **I feel the problem of AI alignment is solvable.**

How?

Alignment research: a history of partial failures

The history of AI safety is littered with failures and partial successes. The most common examples of failure are ideas that would work well typically, but which fail in extreme situations - and a superintelligent AI is perfectly [capable of creating such situations](#).

- [Low-impact AIs](#) were supposed to allow smart machines that interacted with humans without causing huge disruptions. They had some success at 'almost no impact'. But everyone - including me - failed at developing algorithms that had reliable low-impact. If the AI is allowed even a little bit of impact, it can make these low-impact restrictions irrelevant.
- [Corrigibility](#) and [interruptibility](#) were designed to allow AIs to be reprogrammed even when active and powerful. They have good narrow uses, but [aren't a general solution](#): though the AI would not interfere with the interruptibility process, it also has no incentive to preserve it or to ensure its subagents were also interruptible.
- [Oracles](#), question answering AIs (and their close relatives, [tool AIs](#)) are perennial suggestions, the idea being to limit the power of the AI by limiting it to answering questions or giving suggestions. But [that fails](#), for instance when the AI is incentivised to manipulate humans through the contents of its answers or suggestions.
- There were some interesting examples on [limiting AI power](#), but these were ultimately vulnerable to the AI [creating subagents](#).
- The different forms of value learning confronted a [surprising obstacle](#): values could not be learnt without making strong assumptions about human rationality, and human rationality could not be learnt without making strong assumptions about human values.

A litany of partial failures suggests that the next approach tried will be a failure as well - unless we can identify why the approaches above failed. Is there a common failure mode for all of them?

The common thread: lack of value extrapolations

It is easy to point at current examples of agents with low (or high) impact, at safe (or dangerous) suggestions, at low (or high) powered behaviours. So we have in a sense the 'training sets' for defining low-impact/Oracles/low-powered AIs.

It's extending these examples to the general situation that fails: definitions which cleanly divide the training set (whether produced by algorithms or humans) fail to extend to the general situation. Call this the 'value extrapolation problem'^[1], with 'value' interpreted broadly as a categorisation of situations into desirable and undesirable.

Humans turn out to face similar problems. We have broadly defined preferences in familiar situations we have encountered in the world or in fiction. Yet, when confronted with situations far from these, we have to stop and figure out how our values might possibly extend^[2]. Since these human values aren't - yet - defined, we can't directly input them into an algorithm, so AIs that can't solve value extrapolation can't be aligned with human values.

Value extrapolation is thus necessary for AI alignment. It is also [almost sufficient](#), since it allows AIs to draw correct conclusions from imperfectly defined human data. Combined with well grounded basic human values, it will allow the algorithm to extrapolate as well as humans can - better, in fact, using its superhuman abilities.

If that's successful, AIs that value extrapolate and that start aligned, will remain aligned even as they dramatically change the world and confront the unexpected, re-assessing its reward functions when its world-model changes.

Deployment

We think that once humanity builds its first AGI, superintelligence is [likely near](#), leaving little time to develop AI safety at that point. Indeed, it may be necessary that the first AGI start off aligned: we may not have the time or resources to convince its developers to retrofit alignment to it. So we need a way to have alignment deployed throughout the algorithmic world before anyone develops AGI.

To do this, we'll start by offering alignment as a service for more limited AIs. Value extrapolation scales down as well as up: companies value algorithms that won't immediately misbehave in new situations, algorithms that will become conservative and ask for guidance when facing ambiguity.

We will get this service into widespread use (a process that may take some time), and gradually upgrade it to a full alignment process. That will involve drawing on our research and that of others - we will remain strongly engaged with other research groups, providing tools that they can use and incorporating their own results into our service.

We will refine and develop this deployment plan, depending on research results, commercial opportunities, feedback, and suggestions. Contact us in the comments of this post or from our [website](#).

Thanks to LessWrong

I want to thank LessWrong, as a collective entity, for getting us to the point where such a plan seems doable. We'll be posting a lot here, putting out ideas, asking for feedback - if you can continue giving the same quality of response that you always have (and checking that we ourselves haven't go misaligned!), that's all we can ask from you :-)

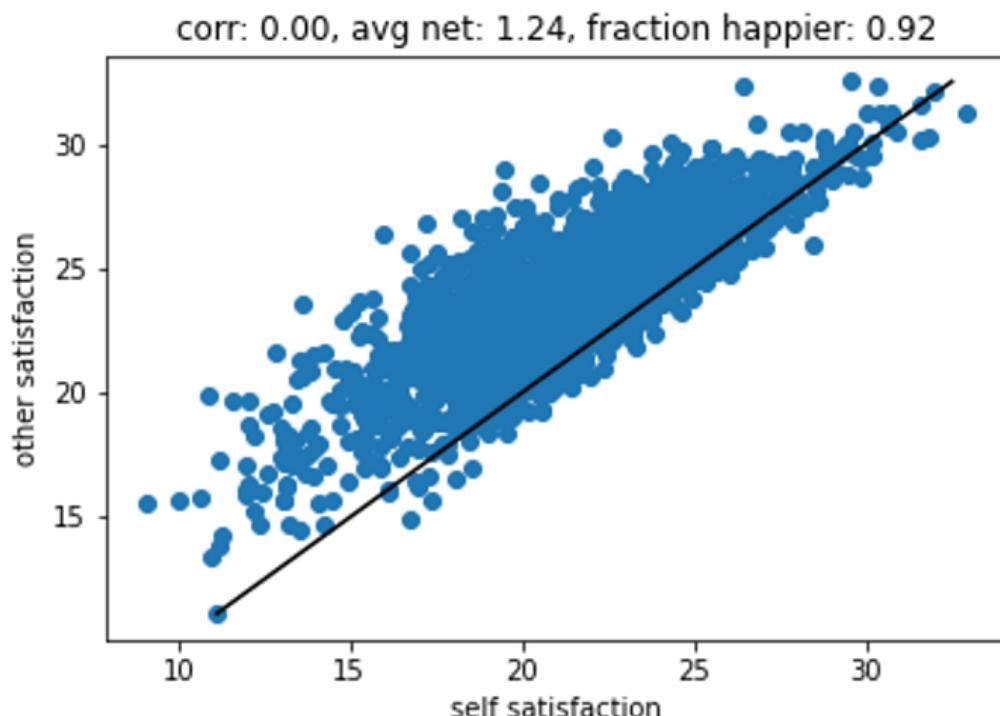
1. Formerly called the 'model splintering' problem. [←](#)
2. Humans have demonstrated a skill with value extrapolation, during their childhoods and adolescences, when encountering new stories and thought-experiments, and when their situation changes dramatically. Though human value extrapolation can be contingent, it rarely falls into the extreme failure modes of AIs. [←](#)

How satisfied should you expect to be with your partner?

I got married today, to the [particular fellow](#) mentioned in [my Turning 30 post](#). In a sort of 'inverse cat tax'^[1] for a sappy announcement, here's a mathematical model of whether you should expect to like your partner more than yourself. I don't mean this in a moralistic way ('thou shalt love thy parents'), tho that might be another post for another time, or necessarily in a utilitarian way ("I would rather they get this ice cream than me"), but as a matter of raw respect ("I think they're a better person than I am, according to my values").

For simplicity's sake, let's consider everyone as having a 'stat' vector and a 'preferences' vector with the same dimensionality, and giving a candidate partner a 'score' based on the dot product of those two vectors. We'll assume that all of the stats are universally good (no one ever prefers an uglier partner over a prettier one, tho they might not care about physical attractiveness much). The preferences vector we'll normalize to have unit magnitude (so it's just an angle in N-dimensional space, basically, defined as a point on the positive sector of the N-spherical shell). For [reasons](#), I'll run simulations with the stat vector as 6 dimensions with 3d6 per stat, leading to a discrete distribution a bit like a truncated normal, with no correlation between the stats.^[2]

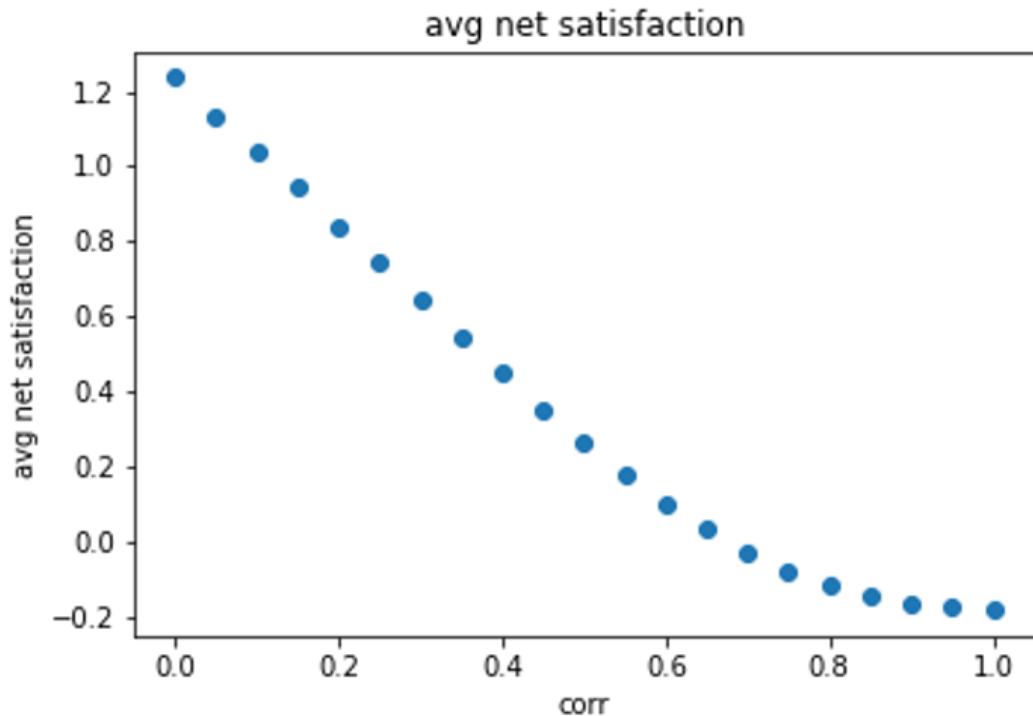
Let's start by considering the heterosexual version of the [stable marriage problem](#), in which people partner up using the well-known Gale-Shapley algorithm, and a simulation with 1,000 each of randomly sampled men and women. Mating is highly assortative; a correlation between total stats of 85.7%, with 83.3% of people have an average total stat difference of less than 6 (the dimensionality).

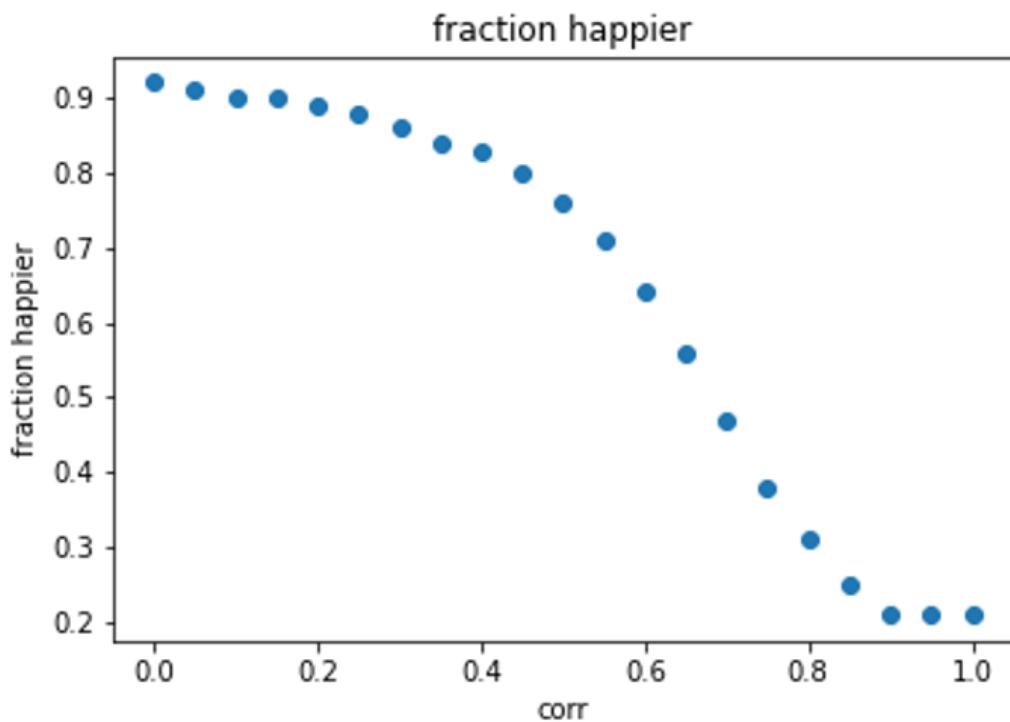


The interesting result is that 91% of people like their partner at least as much as they like themselves, with an average net satisfaction of 1.24. Note that we haven't baked in any

correlation between one's stats and one's preferences, and so this result is, in some sense, not very surprising. The preferences are exerting pressure on the partner (thru who you can stably match with) and not exerting pressure on the self, and so you should expect that pressure to result in higher other-satisfaction.[\[3\]](#)

So let's add an adjustment to the preferences with a scalable parameter corr , so that now it's the (renormalized) sum of the stat vector (times corr) and the previous preference vector (times $\text{abs}(1 - \text{corr})$). As we smoothly vary this from 0 to 1, the average net satisfaction decreases to -0.2 (liking themselves more than their partner) and the fraction happier decreases to 21%.





While the change in net satisfaction is relatively smooth, the change in fraction happier looks much more sigmoidal, with the main drop between $\text{corr} = 0.4$ and $\text{corr} = 0.8$. The main change here is in self-satisfaction, which increases by about 5 points while other-satisfaction increases by only about 3 points.

You can also imagine situations where people specifically want their complements, rather than their mirror. Negative correlations between your stats and your preferences seem unlikely; a more appropriate model seems to be something like relationship satisfaction being a function of the minimum stat between the two partners (or the minimum plus half the maximum, or so on).

The 'marriage' situation with full bisexuality is typically called the [stable roommate problem](#), solved with a similar algorithm. I'll leave it as an exercise for the reader how that impacts the results.^[4]

Anyway, my sense is that when people talk about their 'better half', they're mostly being serious, and this is something that can easily be symmetric.

1. ^

On Imgur, it's common for cat owners to end posts that collect images of use for some other reason with a picture of their cat, referred to as the 'cat tax'.

2. ^

Of course in the real world, everything is correlated; not only is there g for intelligence, but GFP for personality, and wealth causes many material factors to be correlated, and so on. You could try to rationalize this by splitting out the 'natural' variables (like intelligence and wealth) into corrected variables (like intelligence and intelligence-

adjusted wealth), but then it seems odd to have a uniformly random preference vector (as intelligence in the intelligence-adjusted model is more important than in the non-intelligence-adjusted model, given that some wealth-preference has now been moved over to intelligence). I currently don't expect that taking this into account will affect the analysis much (tho doing the analysis with univariate Gaussian stats leads to some odd effects with self-satisfaction, which I'm avoiding here to keep things simple).

3. [^](#)

Correlation between stats and self-satisfaction, of course, is high (0.69), because we insisted that the preferences all be positive, and so people with higher stats will like themselves more accordingly.

4. [^](#)

Naively, I would expect that everyone is more satisfied with their relationships (as they can sample from a wider pool). I think it's likely more assortative in terms of total stats, but it's a little unclear what will happen with the similarity (as corr increases) and what will happen to the crossover point of average net satisfaction (but I'd guess the 0 point is a bit to the right, with the 'increased satisfaction' effect swamping the 'when you try for people similar to you, you can get closer' effect).

Alignment versus AI Alignment

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

Financial status: This work is supported by individual donors and a grant from LTFF.

Epistemic status: This post contains many inside-view stories about the difficulty of alignment.

Thanks to Adam Shimi, John Wentworth, and Rob Miles for comments on this essay.

What exactly is difficult about AI alignment that is not also difficult about alignment of governments, economies, companies, and other non-AI systems? Is it merely that the fast speed of AI makes the AI alignment problem quantitatively more acute than other alignment problems, or are there deeper qualitative differences? Is there a real connection between alignment of AI and non-AI systems at all?

In this essay we attempt to clarify which difficulties of AI alignment show up similarly in non-AI systems, and which do not. Our goal is to provide a frame for importing and exporting insights from and to other fields without losing sight of the difficulties of AI alignment that are unique. Clarifying which difficulties are shared should clarify the difficulties that are truly unusual about AI alignment.

We begin with a series of examples of aligning different kinds of systems, then we seek explanations for the relative difficulty of AI and non-AI alignment.

Alignment in general

In general, we take actions in the world in order to steer the future in a certain direction. One particular approach to steering the future is to take actions that influence the constitutions of some intelligent system in the world. A general property of intelligent systems seems to be that there are interventions one can execute on them that have robustly long-lasting effects, such as changing the genome of a bacterium, or the trade regulations of a market economy. These are the aspects of the respective intelligent systems that persist through time and dictate their equilibrium behavior. In contrast, although plucking a single hair from a human head or adding a single barrel of oil to a market does have an impact on the future, the self-correcting mechanisms of the respective intelligent systems negate rather than propagate such changes.

Furthermore, we will take alignment in general to be about utilizing such interventions on intelligent systems to realize our true terminal values. Therefore we will adopt the following working definition of alignment:

Successfully taking an action that steers the future in the direction of our true terminal values by influencing the part of an intelligent system that dictates its equilibrium behavior.

Our question is: in what ways is the difficulty of alignment of AI systems different from that of non-AI systems?

Example: Aligning an economic society by establishing property rights

Suppose the thing we are trying to align is a human society and that we view that thing as a collection of households and firms making purchasing decisions that maximize their individual utilities. Suppose that we take as a working operationalization of our terminal values the maximization of the sum of individual utilities of the people in the society. Then we might proceed by creating the conditions for the free exchange of goods and services between the households and firms, perhaps by setting up a government that enforces property rights. This is one particular approach to aligning a thing (a human society) with an operationalization of The Good (maximization of the sum of individual utilities). This particular approach works by structuring the environment in which the humans live in such a way that the equilibrium behavior of the society brings about accomplishment of the goal. We have:

- An intelligent system being aligned, which in this case is a human society.
- A model of that system, which in this case is a collection of households and firms making purchasing decisions.
- An operationalization of our terminal values, which in this case is the maximization of the sum of individual utilities (henceforth the "goal").
- A theory – in this case classical microeconomics – that relates possible interventions to the equilibrium behavior of the intelligent system.
- Practical know-how concerning how to establish the conditions that the theory predicts will lead to the goal.

What makes this problem difficult is that it is not so easy to install our intended goal (maximization of the sum of individual utilities) as the explicit goal of the market participants, and even if we did, the participants would not automatically be able to coordinate to achieve that goal without solving the same basic design problem that we are describing here. Similarly, it does not make sense to try to install our goal into the basic computing building-blocks of an AI. The building blocks are mechanical elements that we need to assemble in a way that brings about achievement of the goal as a result of our design choices, and in order to do that we need a theory that connects design choices to outcomes.

The intelligent system may feed a great deal of optimization pressure into the framework that we have placed around it, so if our design choices are even a little bit off the mark then we may not get what we wanted. Furthermore the intelligent system may turn out to have affordances available to it that weren't highlighted by our model of the system, such as when participants in a market economy become actors in the political institutions that regulate the market. If we didn't consider such affordances clearly in our theory then the actual behavior of the system may turn out to be quite different from what we intended.

Finally, the operationalization that we used for our terminal values will generally not be a complete representation of our true terminal values, and, as Goodhart's law suggests, unbounded optimization of any fixed objective generally deviates quite dramatically from the underlying generator that the objective was intended to be a proxy for.

We discuss the differences and similarities between these issues in AI versus non-AI settings below. For now, we turn to further examples of alignment in general.

Example: Aligning a nation by establishing a justice system

Suppose again that the thing we are trying to align is a human society, but now we view this thing as a self-governing polis consisting of individuals with overlapping but not identical social norms. Let us take as our operationalization of our terminal values the existence of the means to peacefully settle disputes. In order to accomplish this we might establish a justice system in which laws must be written down, individuals who are widely respected are appointed as judges and hear cases, prosecutors and defendants get a chance to make their case in public, and clear and binding rulings are made. We can consider the consequences of the choices we make in designing such a system using the tools of game theory, applied ethics, and political science. We might find that certain common-sense design choices would lead to surprising failure modes, while other less obvious design choices lead to equilibria that are closer to our goal. However we approach this, the basic activity is not so different from the problem of aligning an AI. We have:

- An intelligent system being aligned, which in this case is a human society.
- A model of that system, which in this case is a self-governing polis
- An operationalization of our terminal values, which in this case is the fair and predictable enforcement of the law
- A theory – in this case an informal collection of wisdom from game theory, applied ethics, political science, as so on – that relates the possible interventions to the equilibrium behavior of the system
- Practical know-how concerning how to establish the conditions that the theory predicts will lead to the goal

In this case we have the same intelligent system as in the previous example, but a different model of it, which makes possible alignment with respect to a different operationalization of our terminal values using different alignment affordances.

Example: Aligning one's own cognition via habit formation

Suppose that I wish to reduce electricity usage in my house, and that for whatever reason I decide to maintain the following invariant in my house: when there is no-one in a room, the lights in that room should be off. There are various sensors and automated systems I could install to do this, but suppose that the approach I decide on is to establish a personal habit of turning off lights when leaving a room. In order to do that, I might train myself by walking in and out of rooms turning lights off as I leave. Having done that, I create a weekly routine of reviewing all the electricity I've saved and gently thanking myself.

Now if I succeed at firmly establishing this habit then I might affect the pattern of electricity usage in every house I live in over my entire life, and this is actually quite a remarkable feat. Had I installed motion or sound sensors to automatically turn off lights in unused rooms, I would affect electricity usage in my current house but not necessarily in future houses, and the automated system might eventually break or

wear out, and I wouldn't necessarily fix it unless I had installed a habit of doing such things. In contrast, habit formation, if successfully installed, can be long-lived and self-correcting, and this is a general property of steering the future by intervening on intelligent systems.

So here we have:

- An intelligent system being aligned, which is my own cognition.
- A model of that system as a behavioral learning system subject to habit formation.^[1]
- An operationalization of what I care about, which in this case is reducing electricity usage (a much more distant cousin of my true terminal values than the previous examples)
- A theory of habit formation, which in this case is a [behavioral understanding of self-training](#).
- Practical know-how concerning how to execute the training strategy.

Similar to the economic regulation and justice system examples discussed above, it is not so easy to just install our terminal values directly as a single cognitive habit. Perhaps this is possible, but many of us still find reason to systematically install various low-level habits such as thanking friends for their time or getting regular exercise or turning lights off when leaving a room. If it were easy to install our entire terminal values as a single habit then presumably we would do that and have no need for further habit formation.

Also similar to the previous examples, our cognition may exert significant optimization pressure in service of the habits we install, and this may back-fire in the sense of Goodhart's law. We might, for example, deliberately establish a habit of working very hard at our day job, and over time we may as a result be given praise and trusted with further responsibilities, and as a result of this we may come to associate fulfillment of our social needs more and more strongly with the diligence of our work, leading us to even more deeply establish the habit of working hard, leading to further praise and further responsibilities, and so on. There is nothing innately misguided about the habit of working hard at our day job, but the *overall intelligent system* of our cognition may react in powerful and many unexpected ways to the initial establishment of such a habit, leading eventually to actions that no longer serve the original purpose.

Example: Aligning a company via policies and incentives

Suppose I start a company with the objective of building some particular product that I think the world needs. I would like to organize the company in a way that steers the future such that the product comes into existence, and in order to do that I would like to avoid failure mode such as becoming excessively unfocussed, becoming excessively risk-averse, or taking too long to iterate. There are measures that I can take, both before starting the company and while the company is in motion, to steer away from these failure modes. These include formulating and repeating a clear mission statement, setting up a system for promotions that rewards well-calibrated risk taking, and iterating quickly at the beginning of the company in order to habituate a rhythm of quick iteration cycles. Such measures have their effect on the world through aligning the intelligent system that is the company.

So we have:

- An intelligent system being aligned: a collection of people.
- A model of that system as a firm with stakeholders and incentives.
- An operationalization of our terminal values: the product that I think the world needs.
- A theory that relates conditions to consequences: various formal or informal ideas about management, incentivization, legal entity structures, taxation, and so on.
- Practical know-how concerning establishment of the conditions that the theory suggests will lead to the goal

The skill of establishing an effective company lies to a significant extent in honing the skill of aligning intelligent systems with a goal. Company builders can choose to work directly on the object-level problem that the company is facing, and often it is very important for them to do this, but this is because it informs their capacity to align the company as a whole with the goal, and most of their impact on the future flows through alignment. This is why starting a company is so often chosen as the means to accomplish a large goal: humans are the most powerful intelligent system on Earth at the moment, and taking actions that align a group of humans with a particular task is a highly leveraged way to steer the future.

At times it may be most helpful for a founder to view themselves as a kind of Cartesian agent relative to their company, and from that perspective may take actions such as designing an overall reporting structure or identifying bottlenecks as if from "outside the universe". At other times they may view themselves as embedded within the company and may seek to expose themselves to the right information, viewing themselves more as an information processing element that will respond appropriately than like an agent forming plans and taking actions.

Similarly, a founder may at times view the company as a collection of individuals with preferences, as a community with a shared purpose, as a machine operating according to its design principles, or as an information-processing system taking inputs and producing actions as a single entity, and each of these perspectives offers different affordances for alignment. Crucially, however, they all suggest that the way to accomplish a goal is to structure things (policies, incentives, stories, and so forth) in such a way that the overall behavior of an intelligent system (the company) moves the world towards that goal.

Example: Aligning an agentic AI via algorithm design

The classical AI alignment problem is to design an intelligent system that perceives and acts upon the world in a way that is beneficial to all. Suppose that we decide to build an agentic AI that acts in service of an explicit value function, as per early discourse in AI safety, and suppose for concreteness we take as our goal some particular societal objective such as discovering a cure for a particular disease. Here the object being aligned is the AI, and the affordances for alignment are algorithmic modifications to its cognition and value system. By carefully selecting a design for our AI we might hope to steer the future in a very substantial and potentially very precise

way. In order to do this from a pure algorithm-design perspective we will need a theory that connects design choices to the long-term equilibrium behavior of the AI.

Now in constructing an AI from scratch our "affordances" for alignment look less like ways to influence an existing intelligent system and more like design choices for a new intelligent system, and this will be discussed further below under "in-motion versus de-novo alignment". For now we will take this to be a domain with an unusually rich variety of affordances.

So here we have:

- An intelligent system being aligned: the AI.
- A model of that system as an agent executing an algorithm.
- An operationalization of our terminal values: here, the elimination of a disease.
- A (need for a) theory connecting algorithmic design choices to long-run consequences of deploying such an AI.
- Practical know-how concerning algorithm design in light of the path suggested by the theory

In the sections below we will examine the ways that this alignment problem is similar and different to alignment in general, but first we will explore another common framing of the AI alignment problem.

Example: Aligning machine learning systems with training procedures

In another formulation of the AI alignment problem, we take as primary not a space of possible algorithms but a space of possible training methods for various ensembles of machine learning systems. This is often referred to as prosaic AI alignment, and has at times been motivated by the observation that machine learning systems are rapidly becoming more powerful, so we ought to work out how to align them.

In basic machine learning, the affordances for alignment are choices of architecture, initialization, optimization algorithm, and objective function. Beyond this, we can connect multiple learning systems together in ways that check or challenge each other, and we can consider whole algorithms in which the elementary operations consist of training runs. In the framework we are using in this essay, the machine learning approach to alignment begins from a different choice of how to see what an AI is, which naturally suggests different affordances for alignment, just as viewing a human society as a collection of households and firms suggests different affordances for alignment compared to viewing it as a communal story-telling enterprise or as a single giant firm.

Just as in classical AI alignment, we need a [theory that connects design choices to outcomes](#) in order to make intelligent decisions about how to set things up so that our goal will be achieved. In this way, the machine learning approach to alignment is no less theory-driven than the algorithmic approach, though the nature of the theory might be quite different.

So we have:

- An intelligent system being aligned: some combination of machine learning systems
- A model of that system as an optimizer that finds an approximation to a local optima of the objective function
- An operationalization of our terminal values: the training objective.
- A theory connecting training affordances to the goal. A complete theory has obviously proven elusive but we have partial theories in the form of optimization theory, deep double-descent phenomenon, the lottery ticket hypothesis, and so on.
- Practical know-how concerning how to effectively implement the training procedure suggested by the theory.

Example: Aligning domesticated animals through selective breeding

A final example is the selective breeding of a population of domesticated animals as a means to change some trait of interest to us. Suppose for the sake of concreteness that it is a population of dogs we are breeding and hunting ability is the trait we are selecting for. Here the intelligent system is the population as a whole, and the intervention we are making is to select which individuals transmit their genes to the next generation. The gene pool is the thing that determines the "equilibrium behavior" of the population, and our intervention affects that thing in a way that will persist to some extent over time.

One might be tempted to instead say that evolution itself is the thing that we are intervening on, but this seems wrong to us because our intervention does not change the abstract dynamics of evolution, it merely uses evolution to affect a particular population. To intervene on evolution itself would be to reshape the biology of the population so radically that evolution proceeds under fundamentally different dynamics, such by introducing Lamarckian inheritance or asexual reproduction, but this is not what we are considering here.

So we have:

- An intelligent system being aligned: a collection of dogs
- A model of that system as a population subject to natural selection
- An operationalization of our terminal values: the hunting ability of the dogs
- A theory connecting interventions to the goal: the theory of genetics
- Practical know-how concerning how to implement the breeding program

We mention this example because the remainder of our examples are oriented around humans and AIs, and non-human animals represent the main third category.

Non-examples

Here are some examples of things that do not fit the definition of alignment used in this essay.

- Irrigating crops by redirecting a stream. This is not an example of alignment in the sense that we have described here because the stream is not an intelligent system.
- Changing my appearance by getting a haircut. This is not an example of alignment because, although it is an intervention on an intelligent system, it does not really strike at the thing that generates my equilibrium behavior.
- Acquiring water by digging a well. This is not an example of alignment because the action (digging a well) is an object-level task rather than an intervention upon some intelligent system.

In the remainder of this essay we will explore ways in which the difficulty of AI alignment differs from or is similar to that of non-AI systems, with the goal of elucidating a central difference.

Overall risk posed

One axis by which we might differentiate AI and non-AI alignment is the overall level of risk posed on life on Earth by alignment failure. There have been many countries, companies, and communities that were imperfectly aligned with the goals of their designers, but as of the writing of this essay, none of these have ended life on Earth. In contrast, a misaligned AI may destroy all life on the planet. It does seem to us that AI alignment is an outlier in terms of overall risk posed, but why exactly is that? The remainder of this essay explores aspects of AI alignment that make it difficult relative to alignment in general. These might be viewed as explanations for the relatively large risk seemingly posed by AI versus non-AI systems.

Human versus technological speed

When aligning systems composed of humans, there is a match between the speed of the one doing the alignment and that of the system being aligned. If we make a mistake in setting up a government or company, things generally do not run away from us overnight, or if they do, things generally remain under the control of some other human institution if not under our own control. This is because human institutions generally cannot move very much faster than individual humans, which in turn is because the intelligence of a human institution lies significantly within the cognition of individual humans, and we do not yet know how to unpack that.

This match in speed between "aligner" and "alignee" is particularly relevant if the aligner is clarifying their own goals while the "alignee" is forming and executing plans in service of the current goal operationalization. The clarification of "what it is that we really want" seems to be exactly the thing that is most difficult for an aligner to hand off to an aligned, whereas handing off the formulation and execution of plans in service of a particular goal operationalization seems merely very difficult. If we therefore have humans do the clarification of goals while an AI does the formulation and execution of plans then we have two entities that are operating at very different speeds, and we need to take care to get things right.

One approach, then, is to develop the means for slow-thinking humans to oversee fast-thinking AIs without sacrificing on safety, and this is one way to view the work on [approval-directed agents](#) and [informed oversight](#). Another approach is to in fact automate the clarification of goals, and this is one way to view the work on [indirect normativity](#) and [coherent extrapolated volition](#). In the end these two approaches may become the same thing as the former may involve constructing fast imitations of humans that can oversee fast-thinking AIs, which may end up looking much like the latter.

But is this a central difference between AI alignment and alignment in general? It is certainly one difference, but if it were the main difference then we would expect that most of AI alignment would apply equally well to alignment of governments or economies, except that the problem would be less acute in those domains due to the smaller speed difference between aligner and alignee. This may in fact be the case. We will now continue exploring differences.

One-shot versus interactive alignment

When a founder attempts to align a company with a goal, they need not pick a single goal at the outset. Some companies go through major changes of goals, but even among companies that do not, the mission of the company usually gets clarified and adjusted as the company develops. This clarification of goals seems important because unbounded optimization of any fixed operationalization of a goal seems to eventually deviate from the underlying generator of the operationalization as per Goodhart's law. We have not yet found a way to operationalize any goal that does not exhibit this tendency, and so we work with respect to proxies. In the example of turning off the lights in a house this proxy was relatively near-term, while in the microeconomics example of maximizing the sum of individual utilities, this proxy was relatively distant, but in both cases we were working with proxies.

A key issue in AI alignment is that certain AI systems may develop so quickly that we are unable to clarify our goals quickly enough to avoid Goodhart's law. Clarifying our goals means gaining insight from the behavior of the system operating under a crude operationalization, and using that insight to construct a better operationalization, such as when we discover that straightforwardly maximizing the sum of individual revealed preferences in an economy neglects the interests of future humans (it is not that attending to the interests of future humans increases the welfare of current humans but that the welfare of current humans was an incomplete operationalization of what really matters), or when we discover that relentless productivity in our personal lives is depriving us of the space to follow curiosity (it is not that space for curiosity necessarily leads to productivity, but rather than productivity alone was an incomplete operationalization of what really matters).

The phenomenon of an intelligent system outrunning one's capacity to improve the operationalization of one's goals can also happen in fast-growing companies, and it can happen quickly enough that founders fail to make appropriate adjustments. It can also happen in individual lives, for example when one puts in place the conditions to work in a certain field or at a certain job, and these conditions are so effective that one stays in that field or job beyond the point where this is still an effective means to the original end.

There are two basic reactions to this issue in AI alignment: either come up with a goal operationalization that doesn't need to be adjusted, or else make sure one retains the

ability to adjust the goal over time. The former was common in early AI alignment, while the latter is more common now. Corrigibility is a very general formulation of retaining the ability to clarify the goal of an AI system, while [interaction games as formulated at CHAI](#) represent a more specific operationalization. Corrigibility and interactions games are both attempts to avoid a one-time goal specification event.

Is this the central distinguishing feature of AI alignment versus alignment in general? If we did find an operationalization of a goal that never needed to be adjusted then we would certainly have a clear departure from the way that alignment works in other domains. But it seems more likely that if we solve the alignment problem at all, it will involve building AI systems for which we can adjust the goal over time. This is not qualitatively different from setting up a government that can be adjusted over time, or a company that can be adjusted over time, though the problem seems more acute in AI due to the speed at which AI may develop.

Iterative improvement and race dynamics

Human cognition is not something we have the ability to tinker with in the same external way that we can tinker with a toaster oven or jet plane. Companies, economies, and governments are all composed of humans, so we do not have complete access to tinker with everything that's happening in those things, either. We do expect to be able to tinker with AI in the way that we tinker with other technologies, and therefore we expect to be able to make incremental improvements to AI at a rate that is no slower than the general pace of technological improvement. Recognizing this, and recognizing the power that advanced AI systems may open up for their creators, humans may end up in a kind of race to be the first to develop advanced AI systems.

This issue is separate from and additional to the issue of AI systems simply being faster or more capable than humans. It is the expectation of a certain *rate of increase* in AI speed and capabilities that cause race dynamics, since a small head start today could lead to a big advantage later, and it is the *difference* between the rate of AI improvements and the rate of human cognitive improvements that makes such races dangerous, since there is an ever-greater mismatch between the pace at which humans learn by watching the unfolding of a particular intelligent system, and the pace at which those intelligent systems unfold. It is as if we were designing a board game with the goal of making it fun, but the players are AIs that move so quickly that the entire game unfolds before we can learn anything actionable about our game design.

Are race dynamics the fundamental difference between AI and non-AI alignment? It seems to us that race dynamics are more like a symptom of a deeper difference, rather than the central difference itself.

Self-modification

Could it be that the axis that most distinguishes AI alignment from alignment of non-AI systems is a throwback to early discourse on AI alignment: self-modification? Most humans do not seem to deliberately modify themselves to nearly the extent that an AI might be able to. It is not completely clear *why* humans self-modify as little as we do given our wide array of affordances for reshaping our cognition, but whatever the reason, it does seem that AIs may self-modify much more than humans commonly do.

This capacity for self-modification makes AI alignment a challenging technical problem because aligning an entity that considers self-modifying actions requires a strong theory of what it is about that entity that will persist over time. Intuitively, we might construct an agent that acts according to a utility function, and structure its cognition so that it sees that modifying its own values would hinder the achievement of its values. In that way we might establish values that are stable through self-modifying actions. But formulating a theory with which to enact this is a very difficult technical challenge.

Now this problem does come up in other domains. The constitution of the United States has a provision for making constitutional amendments, and this provision could in principle be itself modified by a constitutional amendment. But the constitution of the United States does not have its own agency over the future; it only steers the future via its effect upon a human society, and the humans in that society seem not to self-modify very much.

Conversely, individual humans often do worry about losing something important as we consider self-modifying actions, even though we would seem to have precisely the property that would make self-modification safe for us: namely the ability to reflect on our terminal values and see that changing them would not be in our best interests, since that which is "in our best interests" is precisely that which is aligned with our terminal values.

Now, is this the central difference between AI and non-AI alignment? Self-modification seems like merely one facet of the general phenomenon of embedded agency, yet AIs are certainly not distinguished from non-AI systems by their embeddedness, since all systems everywhere are fundamentally embedded in the physical universe. It seems to us that the seeming strangeness of self-modifying agency is largely an artifact of the relative aversion that humans seem to have to it in their own minds and bodies, and contemporary discourse in AI alignment mostly does not hinge on self-modification as a fundamental distinguishing challenge.

Lack of shared conceptual foundations

Perhaps the reason AI alignment is uniquely difficult among alignment problems is that AI systems do not share a conceptual foundation with humans. When instructing an AI to perform a certain task, the task might be mis-translated into the AI's ontology, or we may fail to include conditions that seem obvious to us. This is not completely different from the way that a written design specification for a product might be mis-understood by a team of humans, or might be implemented without regard for common sense, but the issue in AI is much more acute because there is a much wider gap between a human and an AI than between two humans.

The question, then, is where our terminal values come from, and how they come to us. If they come from outside of us, then we might build AI systems that acquire them directly from the source, and skip over the need to translate them from one set of conceptual foundations to another. If they come from inside of us and are largely or completely shared between people, then we face a translation problem in AI alignment that is very much unique to AI. But probably this very conception of what values are what it would mean for them to "come from" inside or outside of us is confused.

If we manage to clarify the issue of what values are (and whether values are an effective frame for AI alignment in the first place), will we see that the lack of shared conceptual foundations is a central distinguishing feature of AI alignment in comparison to non-AI alignment? Quite possibly. It certainly demands an extreme level of precision in our discourse *about* AI alignment since we are seeking an understanding sufficient for engineering, and such a demand has rarely been placed upon the discourse concerning agency, knowledge, and so forth.

In-motion versus de-novo alignment

When we align our own cognition using habit formation, we are working with an intelligent system that is already in motion, and the affordances available to us are like grasping the steering wheel of a moving vehicle rather than designing a vehicle from scratch. This makes alignment challenging because we must find a way to navigate from where we are to where we want to get in a way that preserves the integrity of our cognition at every point. The same is true, most of the time, when we make changes to the economic, government, and cultural institutions that steer the future via their effect on our society: we are normally working within a society that is already in motion and the affordances available to us consist of making changes on the margin that preserve the integrity of our society at every point.

In AI alignment, one avenue that seems to be available to us is to engineer AI systems from scratch. In this case the "affordances" by which we align an AI with a goal consist of every engineering decision in the construction of the thing, which gives us an exceptional level of flexibility in outcomes. Furthermore, we might do a significant amount of this construction before our AI systems are in motion, which gives us even further flexibility because we are not trying to keep an intelligent system operational during the engineering process.

But when I set up a company, I also have the opportunity to set things up at the outset in order to align the later behavior of the company with my goal. I can design a legal structure, reporting hierarchy, and compensation mechanism before hiring my first employee or accepting my first investment. Some founders do use this opportunity for "de-novo" company engineering to good effect. Similarly, the US constitutional convention of 1787 faced an opportunity for some amount of "de-novo" engineering as the initial constitution of the United States was formulated. Of course, neither of these are truly "de-novo" because the intelligence of the eventual company and nation resides partly or mostly in the internal cognition of the humans that comprise it, and that internal cognition is not subject to design in these examples.

On the other side of the equation, the prosaic AI alignment agenda takes optimization systems as the object of alignment and attempts to align them with a training procedure. This is a kind of mid-way point between alignment of an AI via algorithm design and alignment of a human society by institution design, because the machine learning systems that are taken as primary have more initial structure than the basic elements of algorithm design, but less initial structure than a human society.

Conclusion

The aspects of alignment that we've considered are as follows.

It seems to us that AI alignment as a field is most distinguished from economics, political science, cognitive science, personal habit formation, and other fields concerned with alignment of intelligent systems is that in AI alignment we are forced to get really really precise about what we are talking about, and we are forced to do that all the way up and down the conceptual stack. In contrast, there is only a limited extent to which one really *needs* to understand the basic dynamics of agency when designing economic regulations, or to clarify ethics into really elementary concepts when designing a justice system, or to work out exactly what knowledge is when engaging in personal habit formation.

There are of course fields that attempt to answer such questions precisely, but those fields have not been subject to strong consistent external demands for rigor, and so their level of rigor has been determined mostly by force of will of the participants in those fields. One could view the field of AI alignment as a new high-precision approach to epistemology, metaphysics, and ethics, analogous to the way that the scientific revolution was a new high-precision approach to natural and social inquiry.

To a person living at the beginning of the scientific revolution, it might seem that many great minds had been pouring over the basic questions of natural philosophy for thousands of years, and that little chance therefore existed of making significant contributions to fundamental questions. But from our perspective now it seems that there was a great deal of low-hanging fruit at that time, and it was available to anyone who could summon the patience to look carefully at the world and the courage to test their ideas objectively. The situation we face in AI alignment is different because the low-hanging fruit of empirical investigation have in fact been well-explored. Instead, we are investigating a type of question not amenable to bare empirical investigation, but we are doing so in a way that is motivated by a new kind of demand, and the opportunity for straightforward advances on questions that have eluded philosophers for aeons seems similarly high. The disposition needed is not so much patient observation of natural phenomena but a kind of detail-oriented inquiry into how things must be, coupled with a grounding in something more tangible than that which has guided philosophy-at-large for most of its history.

Just as the laws of thermodynamics were discovered by people working on practical steam engines, and just as both the steam engine and the theory of thermodynamics turned out to be important in the history world, so too the theoretical advances motivated by AI alignment may turn out to be as important as the AI itself. That is, if we don't all die before this field has a chance to flourish. Godspeed.

1. One thing that might make this example confusing is the sense that I "am" my cognition, so the one doing the alignment is the same as the one being aligned. But we don't actually need to take any perspective on such things, because we know from practical experience that it is possible to establish simple habits, and we can see that such habits, if successfully installed, have a kind of flexibility and (potentially) persistence that arise from the intelligence of our cognition. If we like, we can think of ourselves as a kind of "executive / habit machine" in which we are sometimes in habit-formation mode and sometimes in habit-execution mode. ↪

To Change the World

Postdocs are used to disappointment. When Doctor Susan Connor was told she would be taken to the "volcano lair" she thought it was yet another hyperbolic buzzword like "world class", "bleeding edge" and "living wage". She hadn't expected a private jet to fly her to a tropical island complete with a proper stratovolcano.

A regular private jet flight cost as much as Dr Connor earned in a year. If—as Dr Connor suspected—it was a stealth aircraft then that would add an order of magnitude. The VTOL^[1] landed on the short runway. Career academic Susan Connor wasn't used to such white glove treatment but she wasn't complaining either.

Dr Connor was greeted by a tall balding man in a long white labcoat. That broke Dr Connor's credulity. She was a bioinformatician. She had worn a labcoat a handful of times in her entire life—and only when handling toxic materials. This was obviously a psychological experiment. Someone was continuing Stanley Milgram's work. Dr Connor stepped down the airstair as if nothing was amiss.

"Doctor Connor," said the man with his hands spread wide, "I'm Douglas Morbus, Division Chief of the CDP (Center for Disease Proliferation). I enjoyed your recent work on applying entropy-based analysis to junk DNA. It's a pleasure to finally meet you."

"It's nice to meet you Mr Morbus. Or should I call you Dr Morbus?" said Dr Connor.

"Doug, please. We don't bother too much about formalities here, except when welcoming guests of course," said Doug. His freshly-ironed lab coat was bright white. Spotless. Formal attire, apparently, "Full ceremonial dress uniform includes a white fluffy cat but if I brought mine out here she might run off into the jungle and get hurt."

Dr Connor tried to imagine a room full of people with their formal animals. "Hosting a formal ceremony must be like herding cats," said Dr Connor.

"Meetings waste time. We disincentivize them by imposing extraordinary cost," said Doug. He eyed the VTOL.

This was too expensive to be a scientific experiment. Dr Connor was on television. In 2005, a British television station convinced its reality show contestants that they would go into low Earth orbit. (That was many years before the real space tourism industry existed.) They built Russian military base where, for weeks, they taught the contestants fake physics so they wouldn't be surprised at the lack of weightlessness in their fake spaceship.

If this was just a big practical joke then Dr Connor wasn't about to ruin it right away. She wanted to see where it went. Even worse, a part of her wished it was real. Dr Connor wanted to live the harmless supervillain fantasy for just a few minutes longer if that's all it lasted for.

"Follow me," Doug guided Dr Connor down a jungle path, "Effective Evil hires only the best and brightest. We make it easy to get exercise because we want to keep you at peak performance. Hence the network of trails around the island."

Another benefit of the trails would be low production expense. Strolling along a preexisting trail is cheaper than touring a fake laboratory. A free tropical vacation

wasn't a fake spaceship but it was still a free tropical vacation. Dr Connor would take a free tropical vacation over a fake space vacation any day. She'd take a free tropical vacation over a real space vacation too. Space travel sucks.

If the television producers were too cheap to invest in a real fake laboratory then the pranksters would have to earn her compliance in some other way. Dr Connor would test their improvisation.

"I'm curious," said Dr Connor innocently, "Who pays for all this?"

"Is that really the first thing you want to know? Here we are, changing history, and you want to look at our accounting practices? You wouldn't rather hear about all the horrible things we're doing around the world?" said Doug.

Nice try but you're not changing the subject to something you have scripted answers for. "Imagine the funding disappeared. How would I get off the island?" she said.

"We have many escape routes. Aircraft, rockets, ships, submarines. But I'm guessing what you really want to know is if your future funding is secure," said Doug.

Dr Connor nodded. She was a little short of breath. The trail switched back and forth up the volcano.

"Many years ago there was a very rich tech entrepreneur. Founder and CEO of let's-not-talk-about-it," said Doug.

"Was he evil?" said Dr Connor. We are all the heroes of our own story. No real human being would donate money to evil for the sake of evil. Evil is a means to another end. It is not a terminal objective.

"Not at all. There wasn't a selfish bone in his body. He invented cheap medical technologies for developing regions. He'd go undercover in his own company just to check that his employees were being treated well," said Doug.

"Please don't tell me he made a deal with the Devil," said Dr Connor.

"Well...," said Doug.

"I'm sorry. You're telling this story. Go on," said Dr Connor. She hadn't been outside among real living things for a long time. She had forgotten how long it took to get places by foot.

"Philip Goodman put lots of effort into helping other people but not enough into himself. He was obese. It was causing health problems. His doctor told him that if he didn't start exercising he wouldn't live very long," said Doug.

"Oh no," said Dr Connor.

"I mean yeah. He wrote a blockchain contract stipulating that if he didn't put in an average of at least two hours of cardio exercise in every day for a year then his entire fortune would be donated to Effective Evil," said Doug.

Smart contracts are dangerous. Dr Connor didn't know where this was going but it couldn't possibly be good.

"Philip Goodman exercised hard. Harder than he ever had exercised before in his life. His sixty-five-year-old body couldn't handle it. He had a heart attack," said Doug.

Dr Connor gasped.

"That initial investment is what got us started. We still have a steady stream of people who threaten to donate money to Effective Evil unless they accomplish some personal goal—and then they fail—but self-threats are no longer our sole supporters. Government intelligence agencies subsidize us when we destabilize their adversaries. But one of our biggest sources of income is prediction markets. You can make a lot of money from a pandemic prediction market when you're the organization releasing artificial pandemics. We don't do it just for the money, of course, but there's no reason to leave the money lying on the table," said Doug.

"And that's why you need a bioinformatician like me," said Dr Connor.

"That's one of the reasons why we need a bioinformatician," said Doug, "We do human genetic engineering too."

Dr Connor couldn't hold it in any longer. She burst out laughing so hard she nearly lost her balance. She bent over with her hands on her knees until she caught her breath.

"What's so funny?" said Doug with a straight face.

"This whole operation! It's the funniest practical joke anybody has ever played on me," said Dr Connor.

"Dr Connor, I assure you this whole operation is completely legitimate. Well, not legitimate *per se*. We are behind numerous illigitimate activities. But I assure you Effective Evil is completely real," said Doug.

"You're kidding," said Dr Connor.

"I'm not," said Doug.

The forest path ended. They reached a concrete wall laced with barbed wire. Doug scanned them through the checkpoint.

"Most people want to see our breeder reactor first, but in my opinion the hardware related to our cyber-ops is cooler," said Doug.

"The bioengineering facilities please," said Dr Connor. As her area of expertise, it would be the hardest to fake.

It was a real bioengineering facility, complete with mouse cages and cloning vats.

"You're not kidding," said Dr Connor.

"Nope," said Doug.

"What is wrong with you? This is evil," said Dr Connor.

"Thank you," said Doug.

"Why?" said Dr Connor.

"Why what?" said Doug.

Dr Connor gestured frantically at the surrounding facility.

"We're 'Effective Evil', Not 'Theoretical Evil'. I have no patience for the armchair sociopaths who pontificate about villainy without getting their hands dirty," said Doug.

"You're literally wearing lab gloves," said Dr Connor.

"It's a figure of speech," said the supervillain.

Doug escorted the young scientist along the steel catwalk. Chemical engineers labored below.

"I can't join you," said Dr Connor.

"Why not?" said Doug.

"You're evil. With a literal capital 'E'," said Dr Connor.

"So?" said Doug.

"I don't want to make the world a worse place," said Dr Connor.

"You don't have to. There is an oversupply of postdocs. You're a great scientist but (no offence) the marginal difference between hiring you and hiring the next bioinformatician in line is (to us) negligible. Whether or not you (personally) choose to work for us will produce an insignificant net effect on our operations. The impact on your personal finances, however, will be significant. You could easily offset the marginal negative impact of working for us by donating a fraction of your surplus income to altruistic causes instead," said Doug.

"You're proposing I work for you to spread malaria and use my income to subsidize malaria eradication," said Dr Connor.

Doug shrugged. "It's your money," he said.

Dr Connor rested her head on the steel railing. "I think the fumes are getting to my head. Can we go back outside?"

They walked along the forest path back to the VTOL landing pad.

"I became a scientist because I wanted to change the world," said Dr Connor.

"There are no better opportunities to change the world than here at Effective Evil," said Doug.

"I meant 'change the world for the better',' said Dr Connor.

"Then you should have been more specific," said Doug.

Dr Connor stepped back onto the airstair to re-enter the VTOL.

"What about you?" said Dr Connor.

"What about me?" said Doug.

"Why do you run this place?" said Dr Connor.

"Because I want to change the world," said Doug.

1. Vertical take-off and landing [aircraft]. [←](#)

OpenAI Solves (Some) Formal Math Olympiad Problems

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

Epistemic status: I have just skimmed through OpenAI's blogpost and paper, I do not fully understand the details.

From the [blogpost](#)

We built a neural theorem prover for [Lean](#) that learned to solve a variety of challenging high-school olympiad problems, including problems from the [AMC12](#) and [AIME](#) competitions, as well as two problems adapted from the [IMO](#).

[...]

The prover uses a language model to find proofs of formal statements. Each time we find a new proof, we use it as new training data, which improves the neural network and enables it to iteratively find solutions to harder and harder statements.

From the [paper](#)

We explore the use of expert iteration in the context of language modeling applied to formal mathematics. We show that at same compute budget, expert iteration, by which we mean proof search interleaved with learning, dramatically outperforms proof search only. We also observe that when applied to a collection of formal statements of sufficiently varied difficulty, expert iteration is capable of finding and solving a curriculum of increasingly difficult problems, without the need for associated ground-truth proofs. Finally, by applying this expert iteration to a manually curated set of problem statements, we achieve state-of-the-art on the miniF2F benchmark, automatically solving multiple challenging problems drawn from high school olympiads.

Method

- Uses the Lean formal environment instead of the Metamath used in GPT-f.
- Uses "decoder-only Transformers similar to GPT-3" with 774M trainable parameters
- Pre-trained "successively on GPT-3's postprocessed version of CommonCrawl (for 300B tokens) and an updated version of WebMath (for 72B tokens)"
- "proof search interleaved with learning"

The two IMO-adapted problems

Problem 1: Suppose a, b, c are the sides of a triangle. Prove that $a^2(b + c - a) + b^2(c + a - b) + c^2(a + b - c) \leq 3abc$.

Problem 2: For a, b, c reals, prove that $(a^2 + ab + b^2)(b^2 + bc + c^2)(c^2 + ca + a^2) \geq (ab+bc+ca)^3$.

Both solutions to those problems use "nlinarith" applied to the right arguments, which, as far as I understand, is a tactic from mathlib for solving nonlinear arithmetic problems by adding more assumptions to the context of the solver. ([source](#))

The right arguments for the first problem are said in the blogpost to come (informally) from [Schur's inequality](#), which gives

```
nlinarith [sq_nonneg (b - a), sq_nonneg (c - b), sq_nonneg (c - a)]
```

The second problem is solved by applying the Cauchy-Schwarz multiple times, then using some inequality it "invented", and ends up with the same nlinarith expression above.

Related bets and forecasts

- On Metaculus, the question [AI Wins IMO Gold Medal](#) has for community Prediction **Dec 26, 2032** and Metaculus Prediction (different weighting) **Apr 3, 2035**.
- In the comments of [Yudkowsky and Christiano discuss takeoff speeds](#), Christiano [ends up](#) with a 8% chance of "For the 2022, 2023, 2024, or 2025 IMO an AI built before the IMO is able to [get gold]" (see also [this comment](#)).

Conspiracy-proof archeology

This is a cross-post from [Telescopic Turnip](#).



Totally legit fossil of a [3m giant](#), excavated near Cardiff, NY.

There is a lot you can learn about history by asking random Frenchmen about World War Two. Which army contributed the most to the pwnage of Germany? According to [Ifop who conducted the survey in 2014](#), it's the United States (49%), followed by the USSR (23%), and Britain (18%). Now I don't expect these numbers to enlighten your understanding of the Past; the detail that makes them interesting is that Ifop had asked the exact same question three times already: in 2004, in 1994 and in 1945, back when people remembered the war not from textbooks but from echoic memory. Here are the results:

| Survey year | USA | USSR | UK |
|-------------|-----|------|----|
| 1945 | 20 | 57 | 12 |
| 1994 | 49 | 25 | 16 |
| 2004 | 58 | 20 | 16 |
| 2014 | 49 | 23 | 18 |

[Source](#) (in French)

Ask the same question about history several times, and it becomes meta-history. This survey caught live footage of collective memory being overwritten by the victors. Presumably, this happened in a somewhat liberal democracy with a somewhat free press, maybe with a little help from [entertainment](#). It didn't require a totalitarian power deliberately distorting history to manipulate the masses. But the masses were still manipulated somehow.

Of course, for most of history before the quantum revolution, Europe was ruled by totalitarian powers deliberately distorting history to manipulate the masses. If French people's beliefs about the Red Army could drift so easily in the late 20th century, then it's hard to trust anything coming from the official records of divine-right monarchs. You can find historical documents and artifacts and get an idea of what [most likely](#) happened, but if you allow for the possibility that the [power structures](#) of the time could fabricate documents and plant artifacts as much as they wanted, you will never be truly convinced.

Zero-day breaches in the laws of the Universe

In the world of computer hacking, there is a kind of security breach called a [zero-day](#). Zero-days happen when an attacker discovers a breach that the developers themselves are not aware of - it has been known for *zero days*. So there is nothing they can do to prevent an attack. The 2010 [Stuxnet](#) attack on the Iranian nuclear program relied on no less than four different zero-days in Microsoft Windows - not something your typical basement h4xx0r can do. In essence, finding a zero-day boils down to understanding how some part of the computer system works, finding a specific thing that does not work the way everyone else thinks it does, and use it to accomplish something previously thought impossible.

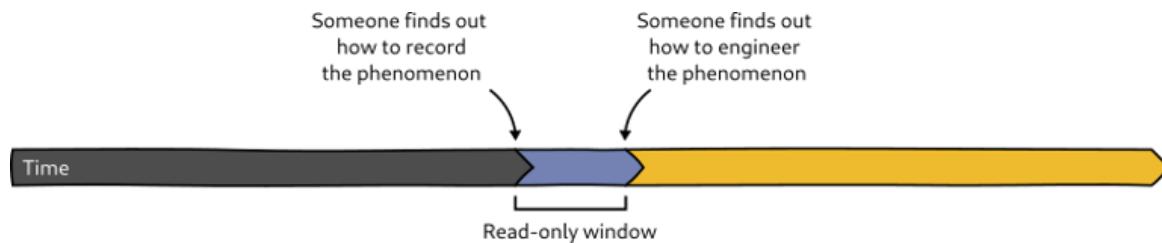
This is not so different from what happens when a scientist discovers a new phenomenon in nature. Everybody thinks the Universe works in a certain way, then someone does an

experiment which clearly deviates from the predictions, and our model of the Universe has to be patched.

Such scientific zero-days make it possible to perform *experimental tests of history*, in a way the most powerful conspiracies and divine-right monarchs could not anticipate^[1]. The only assumption is that the conspiracies/monarchs didn't have access to any futuristic technology or yet-unknown scientific knowledge.

Read-only windows to the past

To test history experimentally, we can exploit the delay between scientific discovery and engineering. First, people discover a phenomenon and come up with ways to measure and record it. Later, people devise new technologies to modify and engineer the phenomenon. Between the two, there is a period that I will call the *read-only window*:



During the ROW, historical evidence can be recorded, but not manipulated. For example, photography was invented in the early 1800s, and the first photomontages were produced [in the mid-1850s](#). So the ROW for photography lasted for about 50 years, after what it became read/write, and of course the powerful used it for political manipulation [throughout the following centuries](#).



This was still considered solid evidence in the 1930s.

Read-only windows are not "on/off", they close gradually. Think of it as the product of the willingness of people to lie, and their capabilities to engineer a fake without getting caught. You can always imagine a secret society powerful enough to fabricate evidence regardless of the year, but the earlier you go back, the more implausibly powerful the secret society needs to be.

How many ROWs are still open today? Let's review some candidates.

Video

Muybridge recorded the classic *Horse in Motion* in 1878, and twenty years later Méliès was already performing all kinds of wizardry on film. The potential for propaganda remained limited until digital video editing [in the 1990s](#). However, the existence of conspiracy theories about the 1969 [moon landing](#) indicates that the ROW for video was already closing back then.

Voice recordings proved harder to fake: the ROW lasted from 1877 (with the first phonographs) to around 2016 (with [voice cloning](#)). As we can now make [fake videos of fake faces with fake voices](#), the read-only windows associated with all these technologies are definitively closed.

Carbon-14

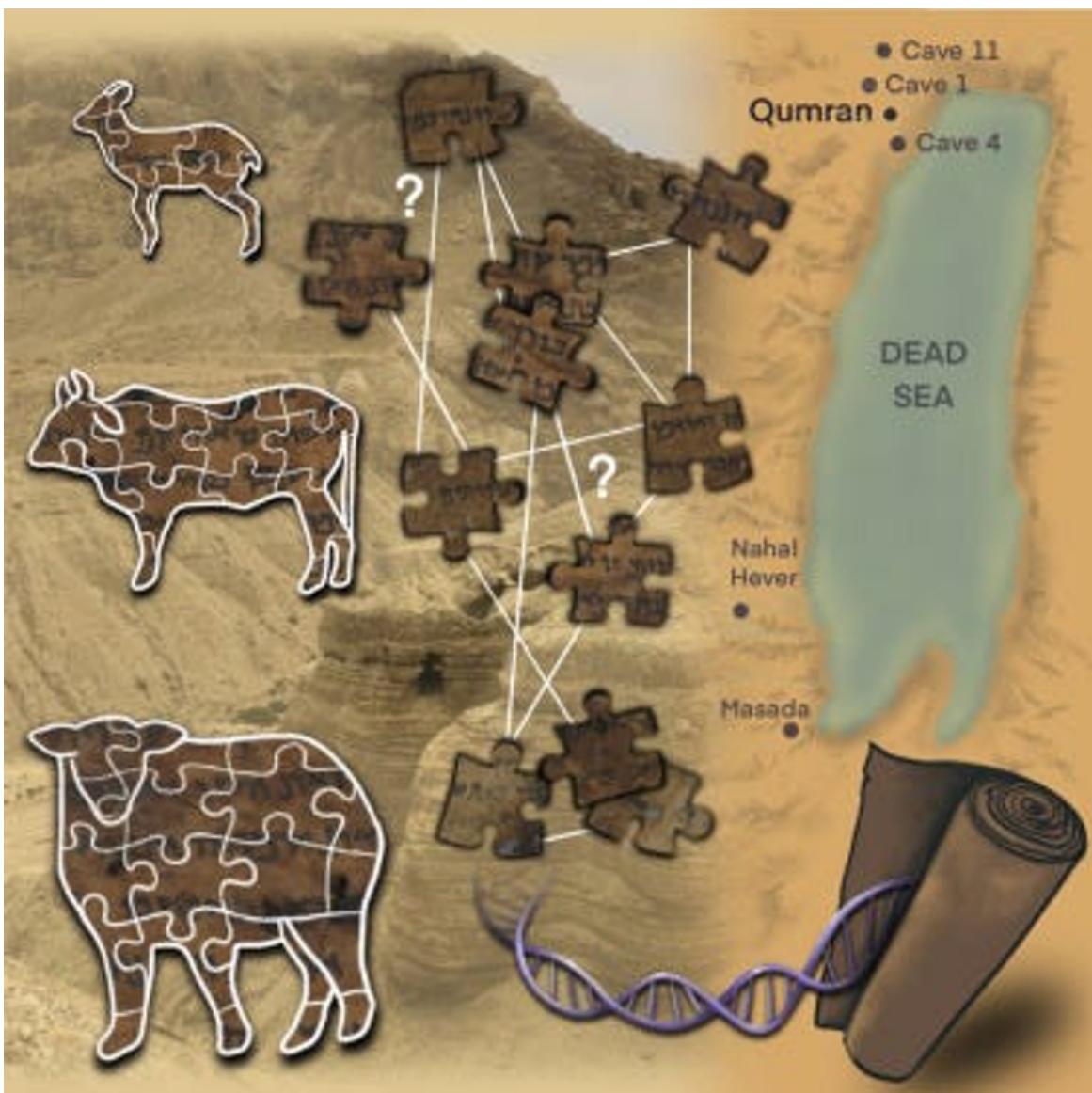
Carbon-14 has changed the official account of history quite a lot. Lately, it was used to show that Vikings were definitely present in North America [as early as year 1021](#). Is this window still open? In my very naive understanding of radiocarbon dating, one could just sprinkle

some carbon mix on various artifacts and plant them in places where they don't belong. If so, this is certainly within the reach of the Secret Viking Misinformation Conspiracy we all believe doesn't exist. If you are a 14C enthusiast, feel free to correct me and explain how actually it's more complicated.

Genetics

The read-only window for genetics opened explosively in the 2000s, when sequencing human genomes went from impossible to "here is another batch of 100,000" in two decades. There are now papers reconstructing the migrations of all kinds of populations ([Japanese](#), [Americans](#), [Inuits](#), [Vikings](#), [French](#), [Middle-Easterners](#), [Africans](#) and of course, [dogs](#)), with some occasionally surprising findings (I was not surprised, because I had no idea what the prior historical theories were anyways).^[2]

But, I can hear you object, none of these uncover any actual juicy conspiracy from the past. Here comes our next source: parchments. Parchments have genomes too, and [there are ways](#) to extract them for sequencing. That's what [Anava et. al](#) did with the [Dead Sea scrolls](#), an ensemble of parchments containing early drafts for the Bible. The whole article is a good read.



Graphical abstract from Anava et al. I'm pretty sure this is from a video game I played in the nineties.

If everything goes the way I hope, this century will see the advent of high-throughput grimoire genomics, and we'll be able to reconstruct the entire goat genealogy to see if it's consistent with the story that is written on the parchments themselves. Perhaps we will discover some sophisticated medieval misinformation, or expose a deep conspiratorial rabbit hole involving the Order of Solomon's Temple, corrupt monastic scribes and, obviously, the Vikings. Science can make this happen.

Is the ROW of genetics still open? Not so fast. Of course it's impossible to fake the ancestry of entire populations, but it's becoming easier and easier to synthesize forged DNA and insert it in individual artifacts. Consider the "lab leak" origin story of SARS-cov-2. One of the main arguments for it is that no direct ancestors of the human-infecting virus have been found so far in animals. Of course, evidence for a lab leak would be a political catastrophe, so the incentives are high for the CCP to avoid it at all costs. If someone finally [found the missing link](#) by sequencing samples from <[any animal from this list](#)>, would you believe it?

It's certainly possible for a large government to covertly design and synthesize a fake missing-link virus to defuse potential lab leak suspicions. Synthesizing a 30 kbp viral genome would cost ~\$10,000 from the [Chinese company SBS](#), and recombinant viruses can be expressed and purified in a few weeks ([here are detailed instructions^{\[3\]}](#)). Making sure the fake missing link can infect both bats and human cells requires more engineering, but nothing impossible. And of course, they would need to do everything clandestinely, with disastrous consequences if the public were to learn about it.

Ultimately, you have to fall back on your *a priori* trust in the Chinese government, and how competent you think they are at conspiring. I have no doubt genetics will teach us a lot of fascinating things about the past, and most of them will be true – but they will not be conspiracy-proof.

Edit: from the comments, Ben brought up [this study](#) about genetic testing revealing that, at some point in the Plantagenet dynasty, someone was not the legitimate heir to the crown. I also found [this recent study](#) of elephant genomics as a way to investigate into ivory trafficking.

Epigenetics

It sounds weird that you could use epigenetics to investigate conspiracies of the past, but it turns out there is something exactly like that. It checks all the boxes:

- A classic historical conspiracy theory: holocaust denial
- A newly-measurable phenomenon: epigenetic markers
- A read-only window: you can't fake epigenetics

The story goes that concentration camp survivors were treated so badly that the stress induced some long-lasting epigenetic modifications in their cells, and these were even [transmitted to their children](#). Holocaust deniers commonly claim that all the official documents are fake, but it's very implausible that the Elders of Zion or whatever could fake *epigenetics*.

Sadly, the actual epigenetic evidence is rather weak, the samples are tiny and I don't think even the most open-minded, well-meaning, charitable holocaust deniers would accept it as proof. But, come on, this illustrates so perfectly what I mean by read-only windows, that I just had to include it. I still hope larger studies will confirm these findings before the window gets closed. But time is running out: [epigenome editing is on its way](#).

Augury

Yes, there is a ROW for augury. [This paper](#) uses changes in migratory birds location to measure past changes in Earth's climate, which I insist technically counts as augury. The major implication is that if we ridiculously fail at controlling global warming and future governments really want to erase all signs that it ever happened, they will have to move around truckloads of birds, just to be sure. You can never be too paranoid when you secretly rule the world.

All of these examples are about history as it was written by the powers of the past. Can we exploit a read-only window to learn about history as it is currently being written by the powers of the present? This will be the topic of the next episode, [the fingerprints of ideology in science](#).

Thanks to Justis for the feedback.

1. ^

Not to be confused with the people who recreate historical situations in the present to see what happens, [which is also a thing](#).

2. ^

Notice how the Viking paper claims that Vikings were actually black-haired, compared to modern Danes. These Vikings are starting to look a bit sus.

3. ^

This is for lentiviruses, which are used all the time in biotech, so the protocols are very streamlined. Making a SARS-cov-2 relative would be more complicated.

QNR prospects are important for AI alignment research

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

Attention conservation notice: This discussion is intended for readers with an interest in prospects for knowledge-rich intelligent systems and potential applications of improved knowledge representations to AI capabilities and alignment. It contains no theorems.

Abstract

Future AI systems will likely use quasilinguistic neural representations (QNRs) to store, share, and apply large bodies of knowledge that include descriptions of the world and human values. Prospects include scalable stores of “ML-native” knowledge that share properties of linguistic and cognitive representations, with implications for AI alignment concerns that include interpretability, value learning, and corrigibility. If QNR-enabled AI systems are indeed likely, then studies of AI alignment should consider the challenges and opportunities they may present.

1. Background

Previous generations of AI typically relied on structured, interpretable, symbolic representations of knowledge; neural ML systems typically rely on opaque, unstructured neural representations. The concept described here differs from both and falls in the broad category of structured neural representations. It is neither fully novel nor widely familiar and well explored.

The term “quasilinguistic neural representations” (QNRs) will be used to denote vector-attributed graphs with quasilinguistic semantics of kinds that (sometimes) make natural language a useful point of reference; a “QNR-enabled system” employs QNRs as a central mechanism for structuring, accumulating, and applying knowledge. QNRs can be language-like in the sense of organizing (generalizations of) NL words through (generalizations of) NL syntax, yet are strictly more expressive, upgrading words to embeddings[1a] (Figure 1) and syntax trees to general graphs (Figure 2). In prospective applications, QNRs would be products of machine learning, shaped by training, not human design. QNRs are not sharply distinguished from constructs already in use, a point in favor of their relevance to real-world prospects.[1b]

Motivations for considering QNR-enabled systems have both descriptive and normative aspects — both *what we should expect* (contributions to AI capabilities in general) and *what we might want* (contributions to AI alignment in particular).[1c] These are discussed in (respectively) Sections 2 and 3.

[1a] For example, embeddings can represent images in ways that would be difficult to capture in words, or even paragraphs (see Figure 1). Embeddings have enormous expressive capacity, yet from a semantic perspective are more computationally tractable than comparable descriptive text or raw images.

[1b] For an extensive discussion of QNRs and prospective applications, see "[QNRs: Toward Language for Intelligent Machines](#)", FHI Technical Report #2021-3, here cited as “QNRs”. A brief introduction can be found here: "[Language for Intelligent Machines: A Prospectus](#)".

[1c] Analogous descriptive and normative considerations are discussed in "[Reframing Superintelligence: Comprehensive AI Services as General Intelligence](#)", FHI Technical Report #2019-1, Section 4.

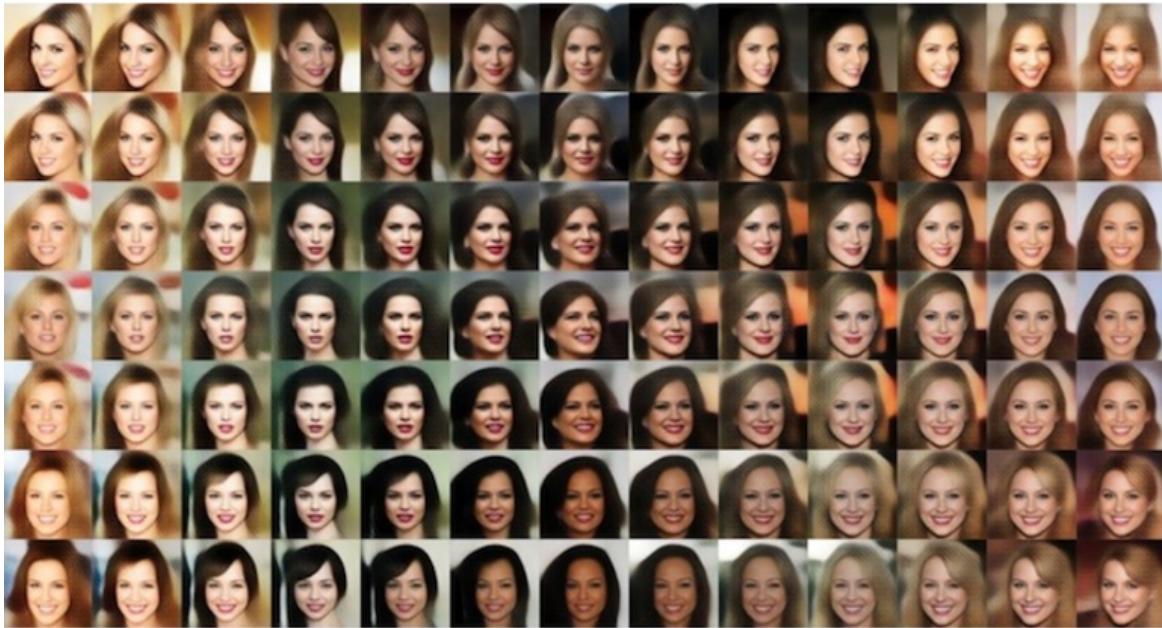


Figure 1: Generalizing semantic embeddings. Images corresponding to points in a two-dimensional grid in a high-dimensional space of face embeddings. Using text to describe faces and their differences in a high-dimensional face-space (typical dimensionalities are on the rough order of 100) would be difficult, and we can expect a similar gap in expressive capacity between embeddings and text in semantic domains where rich denotations cannot be so readily visualized or (of course) described. Image from [Deep Learning with Python](#) (2021).

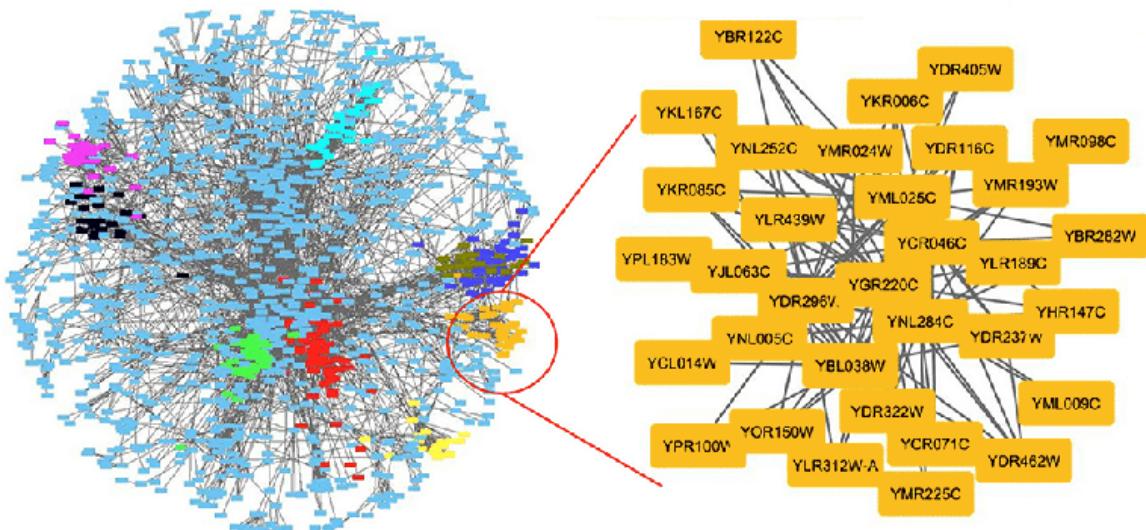


Figure 2: Generalizing semantic graphs. A graph of protein-protein interactions in yeast cells; proteins can usefully be represented by embeddings (see, for example, "[Sequence-based prediction of protein-protein interactions: a structure-aware interpretable deep learning model](#)" (2021)). Image source: "[A Guide to Conquer the Biological Network Era](#)

[Using Graph Theory](#)" (2020). Analogous graphs are perhaps typical of quasilingualistic, ML-native representations of the world, but have a kind of syntax and semantics that strays far from NL. Attaching types or other semantic information to links is natural within a generalized QNR framework.

2. Prospective support for AI capabilities

Multiple perspectives converge to suggest that QNR-enabled implementations of knowledge-rich systems are a likely path for AI development, and taken as a whole can help clarify what QNR-enabled systems might be and do. If QNR-enabled systems are likely, then they are important to problems of AI alignment both as challenges and as solutions. Key aspects include support for efficient scaling, quasi-cognitive content, cumulative learning, semi-formal reasoning, and knowledge comparison, correction, and synthesis.

2.1 Efficient scaling of GPT-like functionality

The cost and performance of language models has increased with scale, for example, from BERT (with 340 million parameters)[2.1a] to GPT-3 (with 175 billion parameters)[2.1b]; the computational cost of a training run on GPT-3 is reportedly in the multi-million-dollar range. Large language models encode not only linguistic skills, but remarkable amounts of detailed factual knowledge, including telephone numbers, email addresses, and the first 824 digits of pi.[2.1c] They are also error-prone and difficult to correct.[2.1d]

The idea that detailed knowledge (for example, of the 824th digit of pi) is best encoded, accurately and efficiently, by gradient descent on a trillion-parameter model is implausible. A natural alternative is to enable retrieval from external stores of knowledge indexed by embeddings and accessed through similarity search, and indeed, recent publications describe Transformer-based systems that access external stores of NL content using embeddings as keys.[2.1e] Considering the complementary capabilities of parametric models and external stores, we can expect to see a growing range of systems in which extensive corpora of knowledge are accessed from external stores, while intensively used skills and commonsense knowledge are embodied in neural models.[2.1f]

...And so we find a natural role for QNR stores (as potential upgrades of NL stores), here viewed from the perspective of state-of-the-art NLP architectures.

[2.1a] "[BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding](#)" (2018).

[2.1b] "[Language Models are Few-Shot Learners](#)" (2020).

[2.1c] "[Extracting training data from large language models](#)" (2021).

[2.1d] Factual accuracy is poor even on simple questions, and it would be beyond challenging to train a stand-alone language model to provide reliable, general, professional-level knowledge that embraced (for example) number theory, organic chemistry, and academic controversies regarding the sociology, economics, politics, philosophies, origins, development, and legacy of the Tang dynasty.

[2.1e] Indexing and retrieving content from Wikipedia is a popular choice. Examples are described in "[REALM: Retrieval-Augmented Language Model Pre-Training](#)," (2020), "[Augmenting Transformers with KNN-Based Composite Memory for Dialog](#)," (2020), and "[Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks](#)" (2021). In a paper last month, "[Improving language models by retrieving from trillions of tokens](#)" (2022), DeepMind described a different, large-corpus-based approach, exhibiting performance

comparable to GPT-3 while using 1/25 as many parameters. In another paper last month, Google reported a system that uses text snippets obtained by an “information retrieval system” that seems similar to Google Search (“[LaMDA: Language Models for Dialog Applications](#)” (2022)).

[2.1f] Current work shows that parametric models and external stores can represent *overlapping semantic content*; stores based on QNRs can deepen this relationship by providing *overlapping semantic representations*. Local QNR structures could correspond closely to graph network states, and standard Transformers in effect operate on fully connected graphs.

2.2 Quasi-cognitive memory

Human memory-stores can be updated by single-shot experiences that include reading journal articles. Our memory-stores include neural representations of things (entities, relationships, procedures...) that are compositional in that they may be composed of multiple parts,[2.2a] and we can retrieve these representations by associative mechanisms. Memories may or may not correspond closely to natural-language expressions — some represent images, actions, or abstractions that one may struggle to articulate. Thus, aspects of human memory include:

- Components with neural representations (much like embeddings)
- Connections among components (in effect, graphs)
- Single-shot learning (in effect, writing representations to a store)
- Retrieval by associative memory (similar to similarity search)[2.2b]

...And so we again find the essential features of QNR stores, here viewed from the perspective of human memory.

[2.2a] Compositionality does not exclude multi-modal representations of concepts, and (in the neurological case) does not imply cortical localization (“[Semantic memory: A review of methods, models, and current challenges](#)” (2020)). Rule representations also show evidence of compositionality (“[Compositionality of Rule Representations in Human Prefrontal Cortex](#)” (2012)). [QNRs](#), Section 4.3, discusses various kinds and aspects of compositionality, a term with different meanings in different fields.

[2.2b] Graphs can be modeled in an associative memory store, but global similarity search is ill-suited to representing connections that bind components together, for example, the components of constructs like sentences or paragraphs. To the extent that connections can be represented by computable relationships among embeddings, the use of *explicit* graph representations can be regarded as a performance optimization.

2.3 Contribution to shared knowledge

To achieve human-like intellectual competence, machines must be *fully literate*, able not only to learn by reading, but to write things worth retaining as contributions to shared knowledge. A natural language for literate machines, however, is unlikely to resemble a natural language for humans. We typically read and write sequences of tokens that represent mouth sounds and imply syntactic structures; a machine-native representation would employ neural embeddings linked by graphs.[2.3a] Embeddings strictly upgrade NL words; graphs strictly upgrade NL syntax. Together, graphs and embeddings strictly upgrade both representational capacity and machine compatibility.

...And so again we find the features of QNR content, here emerging as a natural medium for machines that build and share knowledge.[2.3b]

[2.3a] [QNRs](#), Section 10, discusses potential architectures and training methods for QNR-oriented models, including proposals for learning quasilinguistic representations of high-level abstractions from NL training sets (some of these methods are potentially applicable to training conventional neural models).

[2.3b] Note that this application blurs differences between individual, human-like memory and shared, internet-scale corpora. Similarity search (\approx associative memory) scales to billions of items and beyond; see “[Billion-scale similarity search with GPUs](#)” (2017) and “[Billion-scale Commodity Embedding for E-commerce Recommendation in Alibaba](#)” (2018). Retrieval latency in RETRO (“[Improving language models by retrieving from trillions of tokens](#)” (2022)) is 10 ms.

2.4 Formal and informal reasoning

Research in neurosymbolic reasoning seeks to combine the strengths of structured reasoning with the power of neural computation. In symbolic representations, syntax encodes graphs over token-valued nodes, but neural embeddings are, of course, strictly more expressive than tokens (note that shared nodes in DAGs can represent variables). Indeed, embeddings themselves can express mutual relationships,[2.4a] while reasoning with embeddings can employ neural operations beyond those possible in symbolic systems.

Notions of token-like equality can be generalized to measures of similarity between embeddings, while unbound variables can be generalized to refinable values with partial constraints. A range of symbolic algorithms, including logical inference, have continuous relaxations that operate on graphs and embeddings.[2.4b] These relaxations overlap with pattern recognition and informal reasoning of the sort familiar to humans.

...And so we find a natural role for graphs over embeddings, now as a substrate for quasi-symbolic reasoning.[2.4c]

[2.4a] For example, inference on embeddings can predict edges for knowledge-graph representations; see “[Neuro-symbolic representation learning on biological knowledge graphs](#)” (2017), “[RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space](#)” (2019), and “[Knowledge Graph Embedding for Link Prediction: A Comparative Analysis](#)” (2021).

[2.4b] See, for example, “[Beta Embeddings for Multi-Hop Logical Reasoning in Knowledge Graphs](#)” (2020) and systems discussed in [QNRs](#), Section A1.4.

[2.4c] Transformer-based models have shown impressive capabilities in the symbolic domains of programming and mathematics (see “[Evaluating Large Language Models Trained on Code](#)” (2021) and “[A Neural Network Solves and Generates Mathematics Problems by Program Synthesis](#)” (2022)). As with the overlapping semantic capacities of parametric models and external stores (Section 1, above), the overlapping capabilities of pretrained Transformers and prospective QNR-oriented systems suggest prospects for their compatibility and functional integration. The value of attending to and updating structured memories (perhaps mapped to and from graph neural networks; see “[Graph Neural Networks Meet Neural-Symbolic Computing: A Survey and Perspective](#)” (2021)) presumably increases with the scale and computational depth of semantic content.

2.5 Knowledge accumulation, revision, and synthesis

The performance of current ML systems is challenged by faulty information (in need of recognition and marking or correction) and latent information (where potentially accessible information may be implied — yet not provided — by inputs). These challenges call for comparing semantically related or overlapping units of information, then reasoning about their relationships in order to construct more reliable or complete representations, whether of a thing, a task, a biological process, or a body of scientific theory and observations.[2.5a] This functionality calls for structured representations that support pattern matching, reasoning, revision, synthesis and recording of results for downstream applications.

Relationships among parts are often naturally represented by graphs, while parts themselves are often naturally represented by embeddings, and the resulting structures are natural substrates for the kind of reasoning and pattern matching discussed above. Revised and unified representations can be used in an active reasoning process or stored for future retrieval.[2.5b]

...And so again we find a role for graphs over embeddings, now viewed from the perspective of refining and extending knowledge.

[2.5a] Link completion in knowledge graphs illustrates this kind of process.

[2.5b] For a discussion of potential applications at scale, see [QNRs](#), Section 9. Soft unification enables both pattern recognition and combination; see discussion in [QNRs](#), Section A.1.4.

In light of potential contributions to AI scope and functionality discussed above, it seems *likely* that QNR-enabled capabilities will be widespread in future AI systems, and *unlikely* that QNR functionality will be wholly unavailable. If QNR-enabled capabilities are *likely to be widespread* and relatively easy to develop, then it will be important to consider challenges that may arise from AI development marked by broadly capable, knowledge rich systems. If QNR functionality is *unlikely to be unavailable*, then it will be important to consider how that functionality might help solve problems of AI alignment, in part through differential technology development.

3. Prospective support for AI alignment

Important considerations for AI alignment include interpretability, value learning, and corrigibility in support of strategies for improving behavioral alignment.

3.1 Support for interpretability

In a particularly challenging range of scenarios, AI systems employ opaque representations of knowledge and behaviors that can be understood only through their inputs and outputs. While QNR representations could be opaque, their inherent inductive bias (perhaps intentionally strengthened by training and regularization) should tend to produce relatively compositional, interpretable representations: Embeddings and subgraphs will typically represent semantic units with distinct meanings that are composed into larger units by distinct relationships.[3.1a]

In some applications, QNR expressions could closely track the meanings of NL expressions, [3.1b] making interpretability a matter of lossy QNR → NL translation. In other applications, QNR expressions will be “about something” that can be — at least in outline — explained (diagrammed, demonstrated) in ways accessible to human understanding. In the worst plausible case, QNR expressions will be about recognizable topics (stars, not molecules; humans, not trees), yet substantially opaque in their actual content.[3.1c] Approaches to interpretability that can yield some understanding of opaque neural models seem likely to yield greater understanding when applied to QNR-based systems.

[3.1a] Note that graph edges can carry attributes (types or embeddings), while pairs of embeddings can themselves encode interpretable relationships (as with protein-protein interactions).

[3.1b] For example, QNR semantics could be shaped by NL → NL training tasks that include autoencoding and translation. Interpretable embeddings need not correspond closely to words or phrases: Their meanings may instead correspond to extended NL descriptions, or (stretching the concept of interpretation beyond language *per se*) may correspond to images or other human-comprehensible but non-linguistic representations.

[3.1c] This property (distinguishability of topics) should hold at some level of semantic granularity even in the presence of strong ontological divergence. For a discussion of the general problem, see the discussion of ontology identification in "[Eliciting Latent Knowledge](#)" (2022).

3.2 Support for value learning

Many of the anticipated challenges of aligning agents' actions with human intentions hinge on the anticipated difficulty of learning human preferences. However, systems able to read, interpret, integrate, and generalize from large corpora of human-generated content (history, news, fiction, science fiction, legal codes, court records, philosophy, discussions of AI alignment...) could support the development of richly informed models of human law and ethical principles, together with predictive models of general human concerns and preferences that reflect ambiguities, controversies, partial ordering, and inconsistencies.
[3.2a]

[3.2a] Along lines suggested by Stuart Russell; see discussion in "[Reframing Superintelligence: Comprehensive AI Services as General Intelligence](#)", Section 22.

Adversarial training is also possible: Humans can present hypotheticals and attempt to provoke inappropriate responses; see the use of "adversarial-intent conversations" in "[LaMDA: Language Models for Dialog Applications](#)" (2022).

Training models using human-derived data of the sort outlined above should strongly favor ontological alignment; for example, one could train predictive models of (*human descriptions* of actions and states) → (*human descriptions of human reactions*).
[3.2b] It should go without saying that this approach raises deep but familiar questions regarding the relationship between what people say, what they mean, what they think, what they would think after deeper, better-informed reflection, and so on.

[3.2b] Online sources can provide massive training data of this sort — people enjoy expressing their opinions. Note that this general approach can strongly limit risks of agent-like manipulation of humans during training and application: An automatically curated training set can inform a static but provisional value model for external use.

3.3 Support for corrigibility

Reliance on external, interpretable stores should facilitate corrigibility.
[3.3a] In particular, if distinct entities, concepts, rules, etc., have (more or less) separable, interpretable representations, then identifying and modifying those representations may be practical, a process like (or not entirely unlike) editing a set of statements. In particular, reliance by diverse agents on (portions of) shared, *external stores*
[3.3b] can enable revision by means that are decoupled from the experiences, rewards, etc., of the agents affected. In other words, agents can act based on knowledge accumulated and revised by other sources; to the extent that this knowledge is derived from science, history, sandboxed experimentation, and the like, learning can be safer and more effective than it might be if conducted by (for example) independent RL agents in the wild learning to optimize a general reward function.

[3.3c] Problems of corrigibility should be relatively tractable in agents guided by relatively interpretable, editable, externally-constructed knowledge representations.

[3.3a] “A corrigible agent is one that doesn't interfere with what we would intuitively see as attempts to ‘correct’ the agent, or ‘correct’ our mistakes in building it; and permits these ‘corrections’ despite the apparent instrumentally convergent reasoning saying otherwise.” [“Corrigibility”](#), AI Alignment Forum.

[3.3b] A system can “rely on a store” without constantly consulting it: A neural model can distill QNR content for use in common operations. For an example of this general approach, see the (knowledge graph) → (neural model) training described in [“Symbolic Knowledge Distillation: from General Language Models to Commonsense Models”](#) (2021).

[3.3c] Which seems like a bad idea.

3.4 Support for behavioral alignment

In typical problem-cases for AI alignment, a central difficulty is to provide mechanisms that would enable agents to assess human-relevant aspects of projected outcomes of candidate actions — in other words, mechanisms that would enable agents to take account of human concerns and preferences in choosing among those actions. Expressive, well-informed, corrigible, ontologically aligned models of human values could provide such mechanisms, and the discussion above suggests that QNR-enabled approaches could contribute to their development and application.[3.4a]

[3.4a] Which seems like a good idea.

4. Conclusion

AI systems likely will (or readily could) employ quasilinguistic neural representations as a medium for learning, storing, sharing, reasoning about, refining, and applying knowledge. Attractive features of QNR-enabled systems could include affordances for interpretability and corrigibility with applications to value modeling and behavioral alignment.[4a]

- If QNR-enabled capabilities are indeed *likely*, then they are important to understanding prospective challenges and opportunities for AI alignment, calling for exploration of possible worlds that would include these capabilities.
- If QNR-enabled capabilities are at least *accessible*, then they should be studied as potential solutions to key alignment problems and are potentially attractive targets for differential technology development.

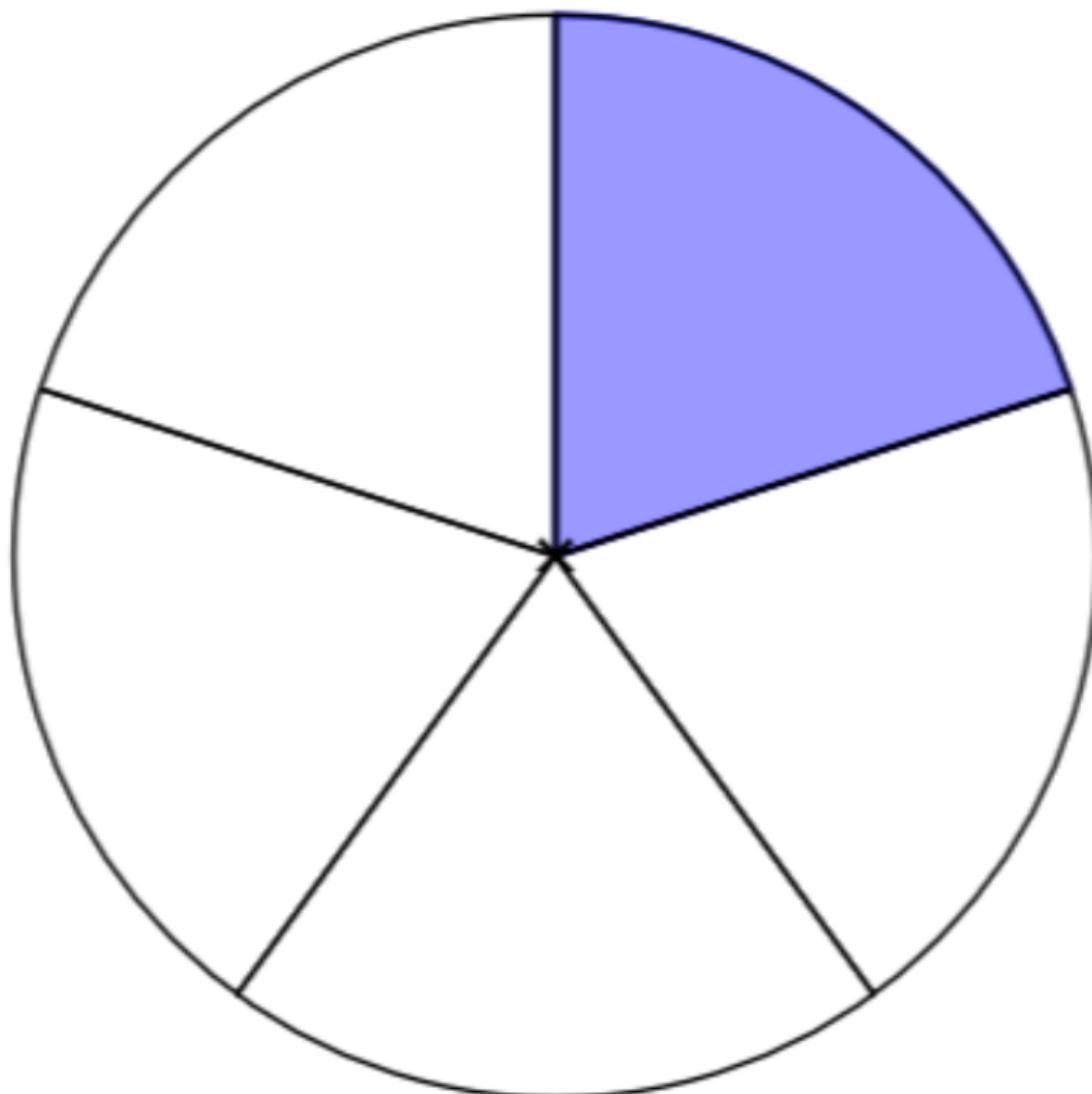
The discussion here is, of course, adjacent to a wide range of deep, complex, and potentially difficult problems, some familiar and others new. Classic AI alignment concerns should be revisited with QNR capabilities in mind.

[4a] Perhaps better approaches will be discovered. Until then, QNR-enabled systems could provide a relatively concrete model of some of what those better approaches might enable.

A Quick Look At 20% Time

I was approached by a client to research the concept of 20% time for engineers, and they graciously agreed to let me share my results. Because this work was tailored to the needs of a specific client, it may have gaps or assumptions that make it a bad 101 post, but in the expectation that it is more useful than not publishing at all, I would like to share it (with client permission).

[Side project time](#), popularized as 20% time at Google, is a policy that allows employees to spend a set percentage of their time on a project of their choice, rather than one directed by management. In practice this can mean a lot of different things, ranging from “spend 20% of your time on whatever you want” to “sure, spend all the free time you want generating more IP for us, as long as your main project is completely unaffected” (often referred to as 120% time) to “theoretically you’re free to do whatever, but we’ve imposed so many restrictions that this means nothing”. I did a 4-hour survey to get a sense of what implementations were available and how they felt for workers.



A frustration here is that almost all of what I could find via Google searches were puff-pieces, anti-puff-pieces, and employees complaining on social media (and one academic article). The [single best article](#) I found came not through a Google search, but because I played D&D with the author 15 years ago and she saw me talking about this on Facebook. She can't be the only one writing about 20% time in a thoughtful way and I'm mad that that writing has been crowded out by work that is, at best, repetitive, and at worst actively misleading.

There are enough anecdotal reports that I believe 20% time exists and is used to good effect by some employees at some companies (including Google) some of the time. The dearth of easily findable information on specific implementations, managerial approaches, trade-offs, etc, makes me downgrade my estimate of how often that happens, vs 20% time being a legible signal of an underlying attitude towards autonomy, or a dubious recruitment tool. I see a real market gap for someone to explain how to do 20% time well at companies of different sizes and product types.

But in the meantime, here's the summary I gave my client. Reminder: this was originally intended for a high-context conversation with someone who was paying me by the hour, and as such is choppier, less nuanced, and has different emphases than ideal for a public blog post.

My full notes are available [here](#).

- To the extent it's measured, utilization appears to be low, so the policy doesn't cost very much.
 - In 2015, a Google HR exec estimated utilization at 10% (meaning it took 2% of all employees' time).
 - In 2009, 12 months after Atlassian introduced 20% time, recorded utilization was at 5% (meaning employees were measured to spend 1.1% of their time on it) and estimated actual utilization was <=15% (Notably, nobody complains that Atlassian 20% is fake, and I confirmed with a recently departed employee that it was still around as of 2020).
- Interaction with management and evaluation is key. A good compromise is to let people spend up to N hours on a project, and require a check-in with management beyond that.
 - Googlers consistently (although not universally) complained on social media that even when 20% time was officially approved, you'd be a fool to use it if you wanted a promotion or raises.
 - However a manager at a less famous company indicated this hadn't been a problem for them, and that people who approached perf the way everyone does at Google would be doomed anyway. So it looks like you can get out of this with culture.
 - An approval process is the kiss of death for a feeling of autonomy, but letting employees work on garbage for 6 months and then holding it against them at review time hurts too.
 - Atlassian requires no approval to start, 3 uninvolved colleagues to vouch for a project to go beyond 5 days, and founder approval at 10 days. This seems to be working okay for them (but see the "costs" section below).
- Costs of 20% time:
 - Time cost appears to be quite low (<5% of employee time, some of which couldn't have been spent on core work anyway)
 - Morale effects can backfire: Sometimes devs make tools or projects that are genuinely useful, but not useful enough to justify expanding or sometimes even maintaining them. This leads to telling developers they must give up on a project they value and enjoyed (bad for their morale) or an abundance of tools that developers value but are too buggy to really rely on (bad for other people's morale). This was specifically called out as a problem at Atlassian.

- Employees on small teams are less likely to feel able to take 20% time, because they see the burden of core work shifting to their co-workers. But being on a small team already increases autonomy, so that may not matter.
- Benefits of 20% time:
 - New products. This appears to work well for companies that make the kind of products software developers are naturally interested in, but not otherwise.
 - The gain in autonomy generally causes the improvements in morale and thus productivity that you'd expect (unless it backfires), but no one has quantified them.
 - Builds slack into the dev pipeline, such that emergencies can be handled without affecting customers.
 - Lets employees try out new teams before jumping ship entirely.
 - Builds cross-team connections that pay off in a number of ways, including testing new teams.
 - Gives developers a valve to overrule bug fixes and feature requests that their boss rejected from the official roadmap.
- There are many things to do with 20% time besides new products.
 - Small internal tools, QOL improvements, etc (but see “costs”).
 - Learning, which can mean classes, playing with new tools, etc.
 - Decreasing technical debt.
 - Non-technical projects, e.g. charity drives.
- Other notes:
 - One person suggested 20% time worked better at Google when it hired dramatically overqualified weirdos to work on mundane tech, and as they started hiring people more suited to the task with less burning desire to be working on something else, utilization and results decreased.
 - 20% or even 120% time has outsized returns for industries that have very high capital costs but minimal marginal costs, such that employees couldn't do them at home. This was a big deal at 3M (a chemical company) and, for the right kind of nerd, big data.

Thanks to the anonymous client for commissioning this research and allowing me to share it, and my [Patreon patrons](#) for funding my writing it up for public consumption.

A comment on Ajeya Cotra's draft report on AI timelines

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

Ajeya Cotra's [draft report on AI timelines](#) is the most useful, comprehensive report about AI timelines I've seen so far. I think the report is a big step towards modeling the fundamental determinants of AI progress. That said, I am skeptical that the arguments provided in the report should reduce our uncertainty about AI timelines by much, even starting from a naive perspective.

To summarize her report, and my (minor) critique,

- Ajeya builds a distribution over how much compute it would take to train a transformative ML model using 2020 algorithms. This compute distribution is the result of combining the distributions under six different biological anchors, and spans over 20 orders of magnitude of compute.
- To calculate when it will be affordable to train a transformative ML model, Ajeya guesses (1) how much hardware price-performance will fall in the coming decades, (2) how much algorithmic progress will happen, and (3) how much large actors (like governments) will be willing to spend on training a transformative AI.
- Ajeya says that she spent most of her time building the 2020 compute distribution, and relatively little time forecasting price-performance, algorithmic progress, and willingness to spend parameters of the model. Therefore, it's not surprising that I'd disagree with her estimates.

However, my main contention is that her estimates for these parameters are just point estimates, meaning that our uncertainty over these parameters doesn't translate into uncertainty in her final bottom-line distribution of when it will become affordable to train a transformative ML model.

- Separately, I think the hardware forecast is too optimistic. She assumes that price-performance will cut in half roughly every 2.5 years over the next 50 years. I think it's more reasonable to expect hardware progress will be slower in the future.
- When making the appropriate adjustments implied by the prior two points, the resulting bottom-line distribution is both more uncertain and later in the future. This barely reduces our uncertainty about AI timelines compared to what many (perhaps most?) EAs/rationalists believed prior to 2020.

Upfront, I definitely am not saying that her report is valueless in light of this rebuttal. In fact, I think Ajeya's report is useful because of how it allows people to build their own models on top of it. I also think the report is admirably transparent, and unusually nuanced. I'm just skeptical that we should draw strong conclusions about when transformative AI will arrive based on her model alone (though future work could be more persuasive).

The critique, elaborated

A summary of how the compute distribution is constructed

Ajeya Cotra's model is highly parameterized compared to simple alternatives that have sometimes been given (like some of [Ray Kurzweil's graphs](#)). It also gives a very wide distribution over what we should expect.

The core of her model is the "2020 training computation requirements distribution", which is a probability distribution over how much computation it would take to train a transformative machine learning model using 2020 algorithms. To build this distribution, she produces six anchor distributions, each rooted in some analogy to biology.

Two anchors are relatively straightforward: the lifetime anchor grounds our estimate of training compute to how much computation humans perform during childhood, and the evolution anchor grounds our estimate of training compute to how much computation evolution performed in the process of evolving humans from the first organisms with neurons.

Three more anchors (the "neural network anchors") ground our estimate of the inference compute of a transformative model in the computation that the human brain uses, plus an adjustment for algorithmic inefficiency. From this estimate, we can derive the number of parameters such a model might have. Using empirical ML results, we can estimate how many data points it will take to train a model with that number of parameters.

Each neural network anchor is distinguished by how it translates "number of data points used during training" into "computation used to train". The short-horizon neural network assumes that each data point will take very little computation, as in the case of language modeling. The medium and long-horizon neural networks use substantially longer "effective horizons", meaning that they estimate it will take much more computation to calculate a single meaningful data point for the model.

The genome anchor grounds our estimate of the number of parameters of a transformative model to the number of parameters in the human genome. Then, it uses a long effective horizon to estimate how much computation it would take to train this model.

By assigning a weight to each of these anchors, we can construct a probability distribution over the total amount of computation it will take to train a transformative AI using 2020 algorithms. Since the lifetime anchor gives substantial probability on amounts of compute that have already been used in real training runs, Ajeya applies a small adjustment to this aggregate distribution.

The final compute distribution is extremely wide, assigning non-negligible probability between 10^{24} FLOPs to 10^{50} FLOPs.

My core contention

To translate the compute distribution into a timeline of when it will become affordable to train transformative AI, Ajeya produces three forecasts for the following three questions:

1. How quickly will computing prices fall in the coming decades?
2. How fast will we make progress in algorithmic efficiency?
3. How much will large actors be willing to spend on training a transformative AI?

Ajeya comments that,

These numbers are much more tentative and unstable than my hypothesis distributions, and I expect many readers to have substantially more information and better intuitions about them than I do; I strongly encourage interested readers to generate their own versions of these forecasts with [this template spreadsheet](#).

Her estimates for each of these parameters are simply numerical guesses, ie. point estimates. That means that **our uncertainty in these guesses is not reflected in the bottom-line distribution for when training transformative AI should become affordable**. But in my opinion, we should be moderately uncertain in each of these parameters.

I think it's inaccurate to convey the final bottom-line probability distribution (the one with a median of about 2052) as representing all of our uncertainty over the variable being modeled. If we were to make some adjustments to the model to account for our uncertainty over these parameters, the bottom-line distribution would get much wider. As a consequence, it's not clear to me that the Bio Anchors model should narrow our uncertainty about AI timelines by an appreciable degree.

To be clear: I'm not claiming that Ajeya makes any claims about Bio Anchors being a final all-thing-considered AI timelines model. Her real claims are quite modest, and she even performs a sensitivity analysis. Still, I think my takeaway from this report is different from how some others have interpreted it.

To give one example of what I see as an inaccurate conclusion being drawn from Ajeya's report, Holden Karnofsky [wrote](#),

Additionally, I think it's worth noting a **couple of high-level points** from Bio Anchors that **don't depend on quite so many estimates and assumptions...**

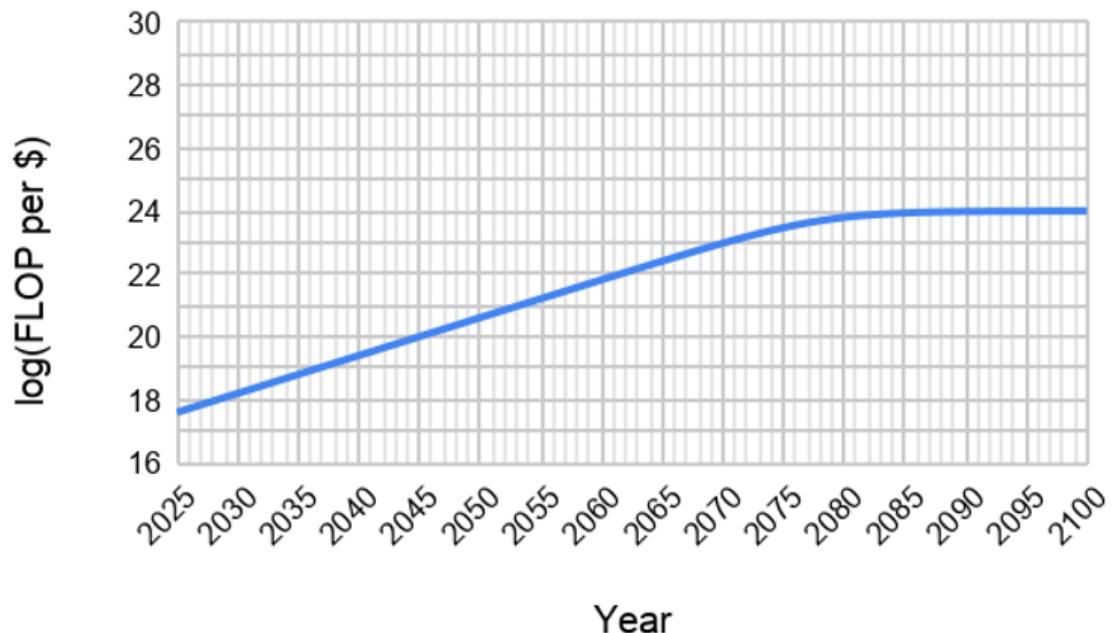
- If AI models continue to become larger and more efficient at the rates that Bio Anchors estimates, it will probably become **affordable this century to hit some pretty extreme milestones - the "high end" of what Bio Anchors thinks might be necessary.** These are hard to summarize, but see the "long horizon neural net" and "evolution anchor" frameworks in the report.
- One way of thinking about this is that the next century will likely see us go from "not enough compute to run a human-sized model at all" to "extremely plentiful compute, as much as even conservative estimates of what we might need."

But, I think the main reason why we hit such high compute milestones by the end of the century in this model is simply that Ajeya assumes that we will have fast progress in computing price-performance in the coming decades; in fact, faster than what we've been used to in the last 10 years. I think this is a dubious assumption.

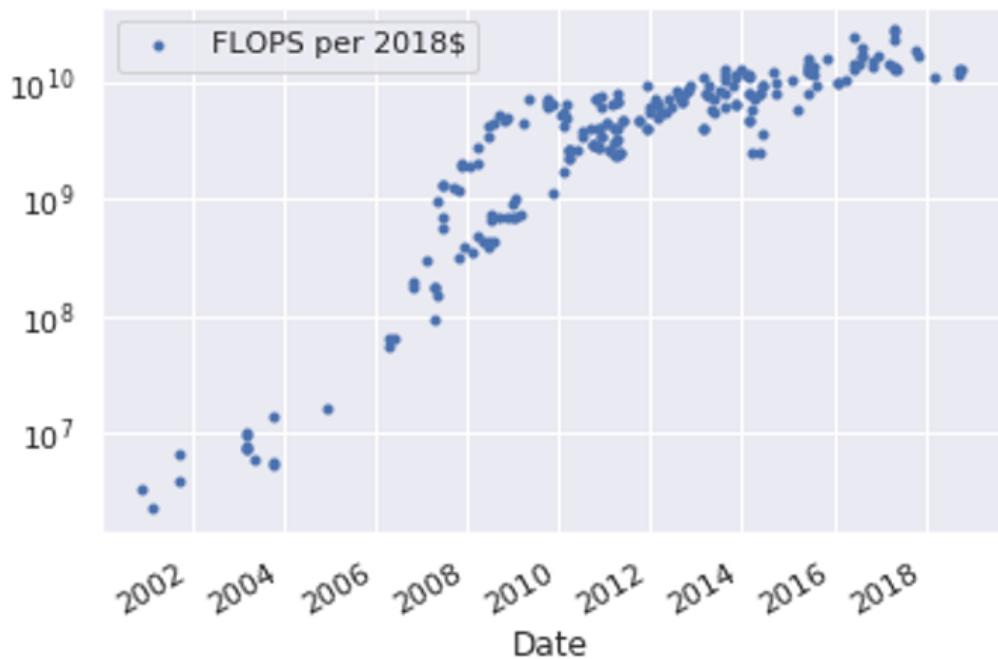
Under [her conservative assumptions from the sensitivity analysis](#), we don't actually get very close to being able to-run evolution by the end of the century.

Her mainline estimate for price-performance progress is that the price of computation will fall by half roughly every 2.5 years for the next 50 years, after which it will saturate. Her mainline forecast looks like this.

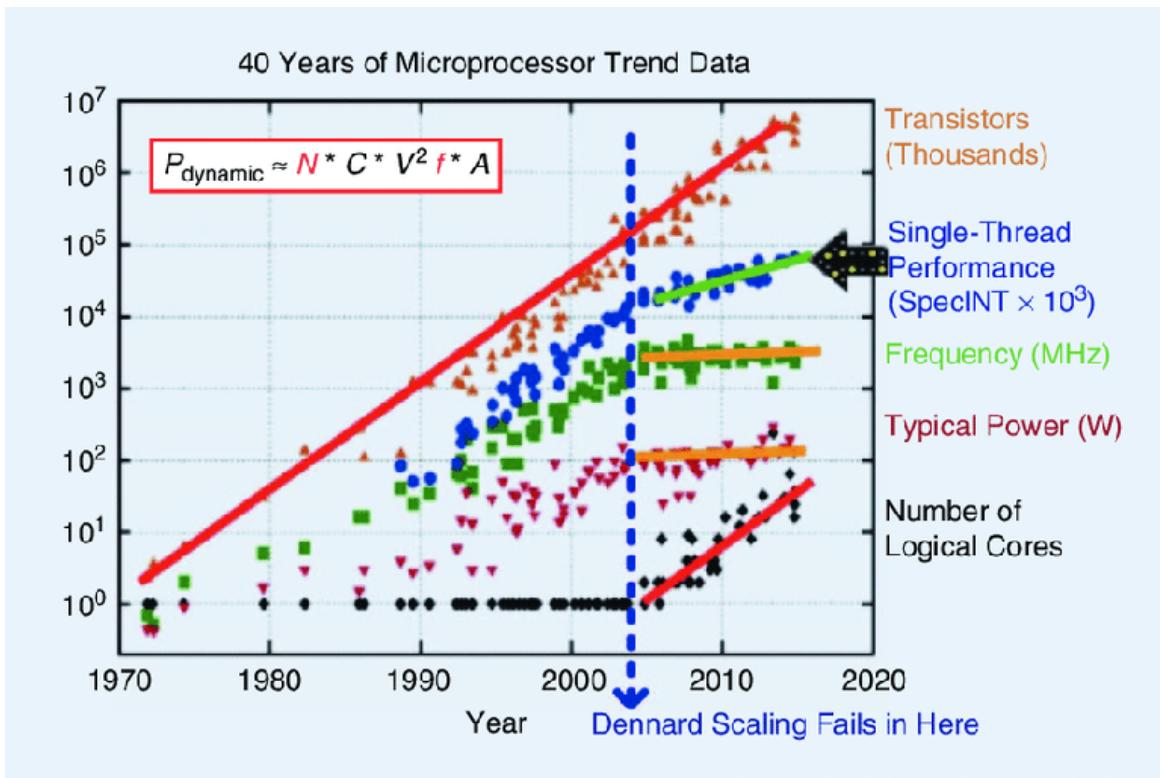
Effective FLOP per dollar by year



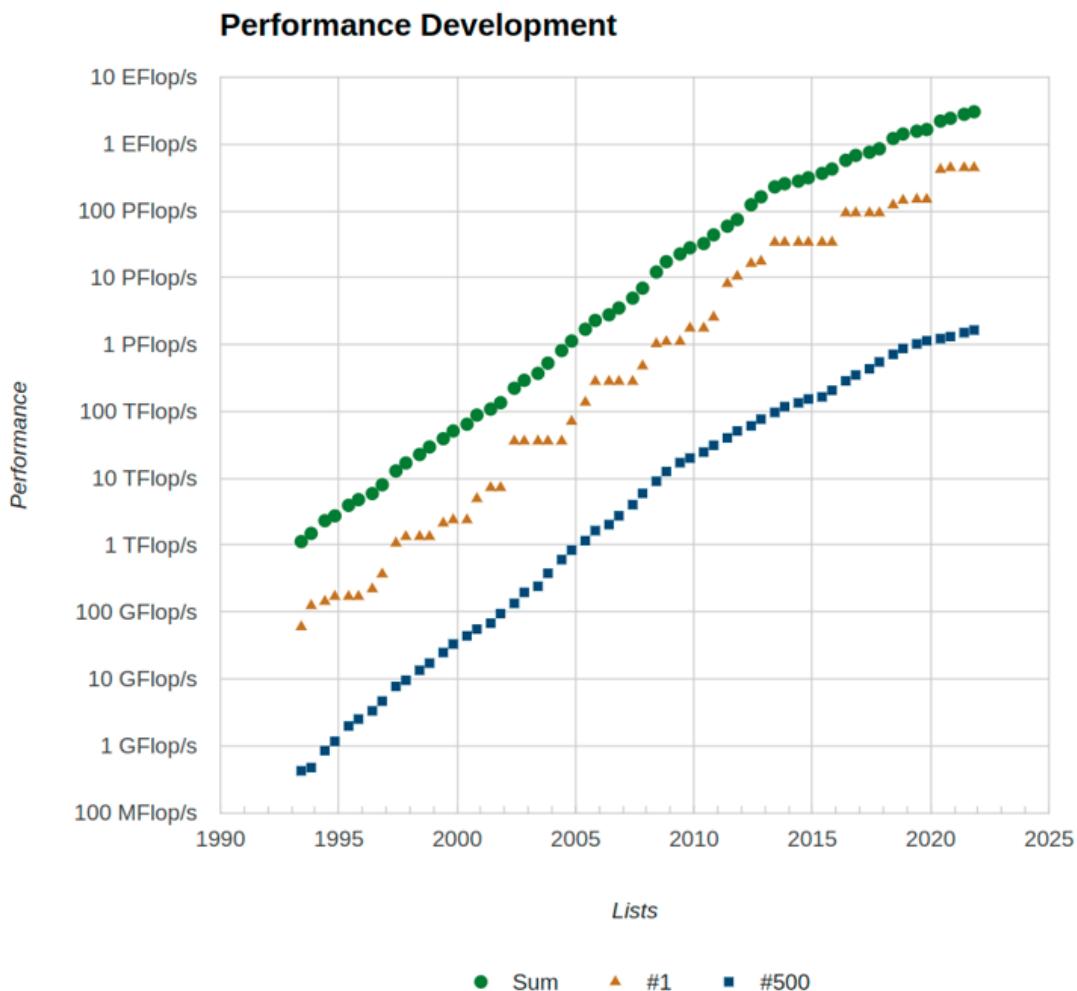
Price performance in the last decade has been quite a lot slower than this. For example, Median Group [found a slower rate](#) of price-performance cutting in half about every 3.5 years in the last 10 years. (I'd also expect a newer analysis to show an even slower trend given the recent [GPU shortage](#)).



Roughly, the main reason for this slowdown was the end of [Dennard scaling](#), which heavily slowed down single-threaded CPU performance.



This slowdown is also in line with a consistent slowdown in supercomputing performance, which is plausibly a good reference class for how much computation will be available for AI projects. Here's a chart from [Top 500](#).



So why does she forecast that prices will cut in half every 2.5 years, instead of a slower rate, matching recent history? In the report, she explains,

This is slower than [Moore's law](#) (which posits a ~1-2 year doubling time and described growth reasonably well until the mid-2000s) but faster than growth in effective FLOP per dollar from ~2008 to 2018 (a doubling time of ~3-4 years). In estimating this rate, I was attempting to weigh the following considerations:

- Other things being equal, the recent slower trend is probably more informative than older data, and is fairly likely to reflect diminishing returns in the silicon chip manufacturing industry.
- However, the older trend of faster growth has held for a much longer period of time and through more than one change in “hardware paradigms.” I don’t think it makes sense to extrapolate the relatively slower growth from 2008 to 2018 over a period of time several times longer than that
- Additionally, a technical advisor informs me that the [NVIDIA A100 GPU](#) (released in 2020) is substantially more powerful than the V100 that it replaced, which could be more consistent with a ~2-2.5 year doubling time than a ~3.5 year doubling time.

Of these points, I find point 2 to be a weak argument in favor of fast future growth, and I find point 3 particularly weak. All things considered, I expect the price-performance trend to stay at its current pace (halving every 3-4 years) or to get even slower, as we get closer to saturating with fundamental physical limits. It just seems hard to imagine that, after experiencing the slowdown we saw in the last decade, hardware engineers will suddenly find a new paradigm that enables much quicker progress in the medium-term.

To get more specific, I'd assign about a 20% credence to the hypothesis that price performance trends will robustly get faster in the next 20 years, and maintain that pace for another 30 years after that. Conversely, that means I think there's about an 80% chance that the hardware milestones predicted in this report are too optimistic.

On this basis, I think it is also reasonable to conclude that we should be similarly conservative for other variables that are downstream from hardware progress, which potentially includes algorithmic progress.

Conclusion

I agree that Ajeya Cotra's report usefully rebuts e.g. Robin Hanson's timelines in which AI is centuries away. However, my impression from talking to rationalists prior to 2020 is that most people didn't put much weight on these sorts of very long timelines anyway.

(Of course, the usual caveats apply. I might just not have been talking to the right people.)

Personally, prior to 2020, I thought AI was very roughly 10 to 100 years away, and this report doesn't substantially update me away from this very uncertain view. That's not to say it's not an interesting, useful report to build on. But it's difficult for me to see how the report, on its own, has any major practical implications for when we should expect to see advanced AI.

Christiano and Yudkowsky on AI predictions and human intelligence

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

This is a transcript of a conversation between Paul Christiano and Eliezer Yudkowsky, with comments by Rohin Shah, Beth Barnes, Richard Ngo, and Holden Karnofsky, continuing the [Late 2021 MIRI Conversations](#).

Color key:

Chat by Paul and Eliezer Other chat

15. October 19 comment

[Yudkowsky][11:01]

thing that struck me as an iota of evidence for Paul over Eliezer:
<https://twitter.com/tamaybes/status/1450514423823560706?s=20>

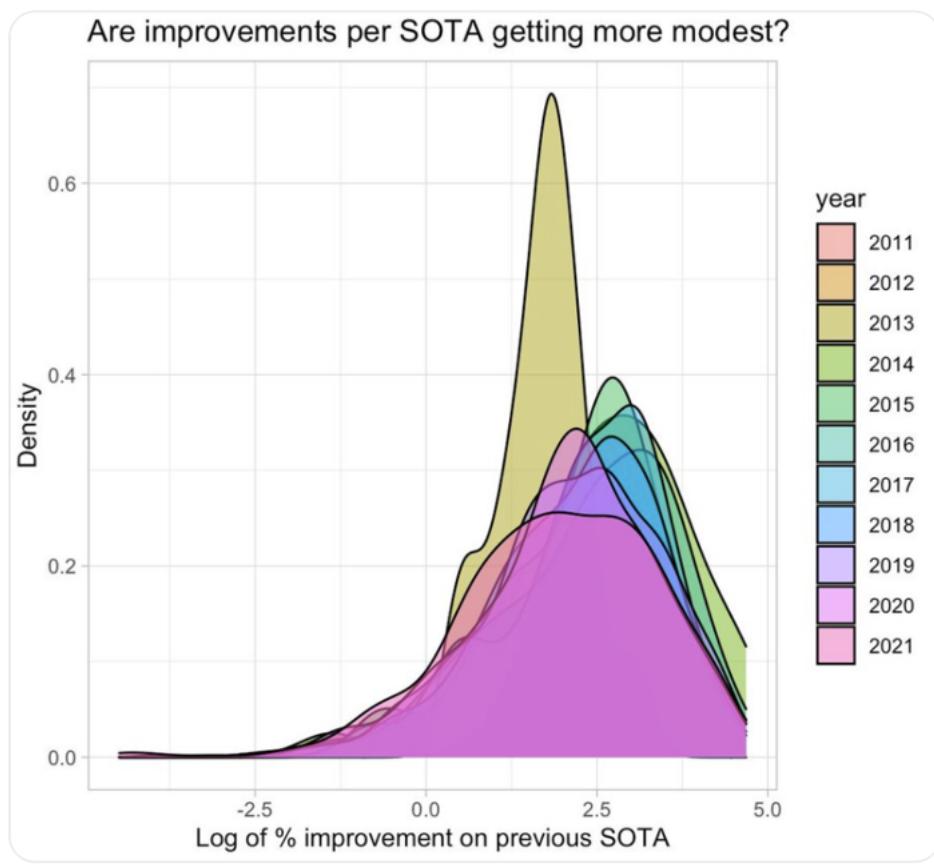


Tamay Besiroglu
@tamaybes

...

Replying to [@tamaybes](#) and [@ESYudkowsky](#)

There is some evidence to suggest that even amongst the top researchers, breakthroughs are getting a little more modest. Here is a density plot showing % improvement on previous SOTA across various years. The data seems to suggest that SOTA improvements get more modest over time.



10:29 AM · Oct 19, 2021 · Twitter Web App

16. November 3 conversation

16.1. EfficientZero

[Yudkowsky][9:30]

Thing that (if true) strikes me as... straight-up falsifying Paul's view as applied to modern-day AI, at the frontier of the most AGI-ish part of it and where Deepmind put in substantial effort on their project? EfficientZero (allegedly) learns Atari in 100,000 frames. Caveat: I'm not having an easy time figuring out how many frames MuZero would've required to achieve the same performance level. MuZero was trained on 200,000,000 frames but reached what looks like an allegedly higher high; the EfficientZero paper compares their performance to MuZero on 100,000 frames, and claims theirs is much better than MuZero given only that many frames.

<https://arxiv.org/pdf/2111.00210.pdf> CC: @paulfchristiano.

(I would further argue that this case is important because it's about the central contemporary model approaching AGI, at least according to Eliezer, rather than any number of random peripheral AI tasks

[Shah][14:46]

I only looked at the front page, so might be misunderstanding, but the front figure says "Our proposed method EfficientZero is 170% and 180% better than the previous SoTA performance in mean and median human normalized score [...] on the Atari 100k benchmark", which does not seem like a huge leap?

Oh, I incorrectly thought that was 1.7x and 1.8x, but it is actually 2.7x and 2.8x, which is a bigger deal (though still feels not crazy to me)

[Yudkowsky][15:28]

the question imo is how many frames the previous SoTA would require to catch up to EfficientZero

(I've tried emailing an author to ask about this, no response yet)

like, perplexity on GPT-3 vs GPT-2 and "losses decreased by blah%" would give you a pretty meaningful concept of how far ahead GPT-3 was from GPT-2, and I think the "2.8x performance" figure in terms of scoring is equally meaningless as a metric of how much EfficientZero improves if any

what you want is a notion like "previous SoTA would have required 10x the samples" or "previous SoTA would have required 5x the computation" to achieve that performance level

[Shah][15:38]

I see. Atari curves are not nearly as nice and stable as GPT curves and often have the problem that they plateau rather than making steady progress with more training time, so that will make these metrics noisier, but it does seem like a reasonable metric to track

(Not that I have recommendations about how to track it; I doubt the authors can easily get these metrics)

[Christiano][18:01]

If you think our views are making such starkly different predictions then I'd be happy to actually state any of them in advance, including e.g. about future ML benchmark results.

I don't think this falsifies my view, and we could continue trying to hash out what my view is but it seems like slow going and I'm inclined to give up.

Relevant questions on my view are things like: is MuZero optimized at all for performance in the tiny-sample regime? (I think not, I don't even think it set SoTA on that task and I haven't seen any evidence). What's the actual rate of improvements since people started studying this benchmark ~2 years ago,

and how much work has gone into it? And I totally agree with your comments that "# of frames" is the natural unit for measuring and that would be the starting point for any discussion.

[Barnes][18:22]

In previous MCTS RL algorithms, the environment model is either given or only trained with reward values, and policies, which cannot provide sufficient training signals due to their scalar nature. The problem is more severe when the reward is sparse or the bootstrapped value is not accurate. The MCTS policy improvement operator heavily relies on the environment model. Thus, it is vital to have an accurate one.

We notice that the output s_{t+1} from the dynamic function G should be the same as s_{t+1} , i.e. the output of the representation function H with input of the next observation o_{t+1} (Fig. 2). This can help to supervise the predicted next state \hat{s}_{t+1} using the actual s_{t+1} , which is a tensor with at least a few hundred dimensions. This provides \hat{s}_{t+1} with much more training signals than the default scalar reward and value.

This seems like a super obvious thing to do and I'm confused why DM didn't already try this. It was definitely being talked about in ~2018

Will ask a DM friend about it

[Yudkowsky][22:45]

I... don't think I want to take *all* of the blame for misunderstanding Paul's views; I think I also want to complain at least a little that Paul spends an insufficient quantity of time pointing at extremely concrete specific possibilities, especially real ones, and saying how they do or don't fit into the scheme.

Am I rephrasing correctly that, in this case, if Efficient Zero was actually a huge (3x? 5x? 10x?) jump RL sample efficiency over previous SOTA, measured in 1 / frames required to train to a performance level, then that means the Paul view *doesn't* apply to the present world; but this could be because MuZero wasn't the real previous SOTA, or maybe because nobody really worked on pushing out this benchmark for 2 years and therefore on the Paul view it's fine for there to still be huge jumps? In other words, this is something Paul's worldview has to either deny or excuse, and not just, "well, sure, why wouldn't it do that, you have misunderstood which kinds of AI-related events Paul is even trying to talk about"?

In the case where, "yes it's a big jump and that shouldn't happen later, but it could happen now because it turned out nobody worked hard on pushing past MuZero over the last 2 years", I wish to register that my view permits it to be the case that, when the world begins to end, the frontier that enters into AG is similarly something that not a lot of people spent a huge effort on since a previous prototype from 2 years earlier. It's just not very surprising to me if the future looks a lot like the past, or if human civilization neglects to invest a ton of effort in a research frontier.

Gwern guesses that getting to EfficientZero's performance level would require around 4x the samples for MuZero-Reanalyze (the more efficient version of MuZero which replayed past frames), which is also apparently the only version of MuZero the paper's authors were considering in the first place - without replays, MuZero requires 20 billion frames to achieve its performance, not the figure of 200 million. <https://www.lesswrong.com/posts/jYNT3Qihn2aAYaaPb/efficientzero-human-ale-sample-efficiency-with-muzero-self?commentId=JEHPQa7i8Qjcg7TW6>

17. November 4 conversation

17.1. EfficientZero (continued)

[Christiano][7:42]

I think it's possible the biggest misunderstanding is that you somehow think of my view as a "scheme" and your view as a normal view where probability distributions over things happen.

Concretely, this is a paper that adds a few techniques to improve over MuZero in a domain that (it appears) wasn't a significant focus of MuZero. I don't know how much it improves but I can believe gwern's estimates of 4x.

I'd guess MuZero itself is a 2x improvement over the baseline from a year ago, which was maybe a 4x improvement over the algorithm from a year before that.

If that's right, then no it's not mindblowing on my view to have 4x progress one year, 2x progress the next, and 4x progress the next.

If other algorithms were better than MuZero, then the 2019-2020 progress would be >2x and the 2020-2021 progress would be <4x.

I think it's probably >4x sample efficiency though (I don't totally buy gwern's estimate there), which makes it at least possibly surprising.

But it's never going to be that surprising. It's a benchmark that people have been working on for a few years that has been seeing relatively rapid improvement over that whole period.

The main innovation is how quickly you can learn to predict future frames of Atari games, which has tiny economic relevance and calling it the most AGI-ish direction seems like it's a very Eliezer-ish view, this isn't the kind of domain where I'm either most surprised to see rapid progress at all nor is the kind of thing that seems like a key update re: transformative AI

yeah, SoTA in late 2020 was SPR, published by a much smaller academic group:
<https://arxiv.org/pdf/2007.05929.pdf>

MuZero wasn't even setting sota on this task at the time it was published

my "schemes" are that (i) if a bunch of people are trying on a domain and making steady slow progress, I'm surprised to see giant jumps and I don't expect most absolute progress to occur in such jumps, (ii) if a domain is worth a lot of \$, generally a bunch of people will be trying. Those aren't claims about what is always true, they are claims about what is typically true and hence what I'm guessing will be true for transformative AI.

Maybe you think those things aren't even good general predictions, and that I don't have long enough tails in my distributions or whatever. But in that case it seems we can settle it quickly by prediction.

I think this result is probably significant (>30% absolute improvement) + faster-than-trend (>50% faster than previous increment) progress relative to prior trend on 8 of the 27 atari games (from table 1, treating SimPL->{max of MuZero, SPR}->EfficientZero as 3 equally spaced datapoints): Asterix, Breakout, almost ChopperCMD, almost CrazyClimber, Gopher, Kung Fu Master, Pong, Qbert, SeaQuest. My guess is that they thought a lot about a few of those games in particular because they are very influential on the mean/median. Note that this paper is a giant grab bag and that simply stapling together the prior methods would have already been a significant improvement over prior SoTA. (ETA: I don't think saying "its only 8 of 27 games" is an update against it being big progress or anything. I do think saying "stapling together 2 previous methods without any complementarity at all would already have significantly beaten SoTA" is fairly good evidence that it's not a hard-to-beat SoTA.)

and even fewer people working on the ultra-low-sample extremely-low-dimensional DM control environments (this is the subset of problems where the state space is 4 dimensions,

people are just not trying to publish great results on cartpole), so I think the most surprising contribution is the atari stuff

OK, I now also understand what the result is I think?

I think the quick summary is: the prior SoTA is SPR, which learns to predict the domain and then does Q-learning. MuZero instead learns to predict the domain and does MCTS, but it predicts the domain in a slightly less sophisticated way than SPR (basically just predicts rewards, whereas SPR predicts all of the agent's latent state in order to get more signal from each frame). If you combine MCTS with more sophisticated prediction, you do better.

I think if you told me that DeepMind put in significant effort in 2020 (say, at least as much post-MuZero effort as the new paper?) trying to get great sample efficiency on the easy-exploration atari games, and failed to make significant progress, then I'm surprised.

I don't think that would "falsify" my view, but it would be an update against? Like maybe if DM put in that much effort I'd maybe have given only a 10-20% probability to a new project of similar size putting in that much effort making big progress, and even conditioned on big progress this is still >>median (ETA: and if DeepMind put in much more effort I'd be more surprised than 10-20% by big progress from the new project)

Without DM putting in much effort, it's significantly less surprising and I'll instead be comparing to the other academic efforts. But it's just not surprising that you can beat them if you are willing to put in the effort to reimplement MCTS and they aren't, and that's a step that is straightforwardly going to improve performance.

(not sure if that's the situation)

And then to see how significant updates against are, you have to actually contrast them with all the updates in the other direction where people *don't* crush previous benchmark results and instead just make modest progress

I would guess that if you had talked to an academic about this question (what happens if you combine SPR+MCTS) they would have predicted significant wins in sample efficiency (at the expense of compute efficiency) and cited the difficulty of implementing MuZero compared to any of the academic results. That's another way I could be somewhat surprised (or if there were academics with MuZero-quality MCTS implementations working on this problem, and they somehow didn't set SoTA, then I'm even more surprised). But I'm not sure if you'll trust any of those judgments in hindsight.

Repeating the main point :

I don't really think a 4x jump over 1 year is something I have to "defy or excuse", it's something that I think becomes more or less likely depending on facts about the world, like (i) how fast was previous progress, (ii) how many people were working on previous projects and how targeted were they at this metric, (iii) how many people are working in this project and how targeted was it at this metric

it becomes continuously less likely as those parameters move in the obvious directions

it never becomes 0 probability, and you just can't win that much by citing isolated events that I'd give say a 10% probability to, unless you actually say something about how you are giving >10% probabilities to those events without losing a bunch of probability mass on what I see as the 90% of boring stuff

[Ngo: ]

and then separately I have a view about lots of people working on important problems, which doesn't say anything about this case

(I actually don't think this event is as low as 10%, though it depends on what background facts about the project you are conditioning on---obviously I gave <<10% probability to someone publishing this particular result, but something like "what fraction of progress in this field would come down to jumps like this" or whatever is probably >10% until you tell me that DeepMind actually cared enough to have already tried)

[Ngo][8:48]

I expect Eliezer to say something like: DeepMind believes that both improving RL sample efficiency, and benchmarking progress on games like Atari, are important parts of the path towards AGI. So insofar as your model predicts that smooth progress will be caused by people working directly towards AGI, DeepMind not putting effort into this is a hit to that model. Thoughts?

[Christiano][9:06]

I don't think that learning these Atari games in 2 hours is a very interesting benchmark even for deep RL sample efficiency, and it's totally unrelated to the way in which humans learn such games quickly. It seems pretty likely totally plausible (50%) to me that DeepMind feels the same way, and then the question is about other random considerations like how they are making some PR calculation.

[Ngo][9:18]

If Atari is not a very interesting benchmark, then why did DeepMind put a bunch of effort into making Agent57 and applying MuZero to Atari?

Also, most of the effort they've spent on games in general has been on methods very unlike the way humans learn those games, so that doesn't seem like a likely reason for them to overlook these methods for increasing sample efficiency.

[Shah][9:32]

It seems pretty likely totally plausible (50%) to me that DeepMind feels the same way, and then the question is about other random considerations like how they are making some PR calculation.

Not sure of the exact claim, but DeepMind is big enough and diverse enough that I'm pretty confident at least some people working on relevant problems don't feel the same way

[...] This seems like a super obvious thing to do and I'm confused why DM didn't already try this. It was definitely being talked about in ~2018

Speculating without my DM hat on: maybe it kills performance in board games, and they want one algorithm for all settings?

[Christiano][10:29]

Atari games in the tiny sample regime are a different beast

there are just a lot of problems you can state about Atari some of which are more or less interesting (e.g. jointly learning to play 57 Atari games is a more interesting problem than learning how to play one of them absurdly quickly, and there are like 10 other problems about Atari that are more interesting than this one)

That said, Agent57 also doesn't seem interesting except that it's an old task people kind of care about. I don't know about the take within DeepMind but outside I don't think anyone would care about it other than historical significance of the benchmark / obviously-not-cherrypickedness of the problem.

I'm sure that some people at DeepMind care about getting the super low sample complexity regime. I don't think that really tells you how large the DeepMind effort is compared to some random academics who care about it.

[Shah: ]

I think the argument for working on deep RL is fine and can be based on an analogy with humans while you aren't good at the task. Then once you are aiming for crazy superhuman

performance on Atari games you naturally start asking "what are we doing here and why are we still working on atari games?"

[Ngo: ]

and correspondingly they are a smaller and smaller slice of DeepMind's work over time

[Ngo: ]

(e.g. Agent57 and MuZero are the only DeepMind blog posts about Atari in the last 4 years, it's not the main focus of MuZero and I don't think Agent57 is a very big DM project)

Reaching this level of performance in Atari games is largely about learning perception, and doing that from 100k frames of an Atari game just doesn't seem very analogous to anything humans do or that is economically relevant from any perspective. I totally agree some people are into it, but I'm totally not surprised if it's not going to be a big DeepMind project.

[Yudkowsky][10:51]

would you agree it's a load-bearing assumption of your worldview - where I also freely admit to having a worldview/scheme, this is not meant to be a prejudicial term at all - that the line of research which leads into world-shaking AGI must be in the mainstream and not in a weird corner where a few months earlier there were more profitable other ways of doing all the things that weird corner did?

eg, the tech line leading into world-shaking AGI must be at the profitable forefront of non-world-shaking tasks. as otherwise, afaict, your worldview permits that if counterfactually we were in the Paul-forbidden case where the immediate precursor to AGI was something like EfficientZero (whose motivation had been beating an old SOTA metric rather than, say, market-beating self-driving cars), there might be huge capability leaps there just as EfficientZero represents a large leap, because there wouldn't have been tons of investment in that line.

[Christiano][10:54]

Something like that is definitely a load-bearing assumption

Like there's a spectrum with e.g. EfficientZero --> 2016 language modeling --> 2014 computer vision --> 2021 language modeling --> 2021 computer vision, and I think everything anywhere close to transformative AI will be way way off the right end of that spectrum

But I think quantitatively the things you are saying don't seem quite right to me. Suppose that MuZero wasn't the best way to do anything economically relevant, but it was within a factor of 4 on sample efficiency for doing tasks that people care about. That's already going to be enough to make tons of people extremely excited.

So yes, I'm saying that anything leading to transformative AI is "in the mainstream" in the sense that it has more work on it than 2021 language models.

But not necessarily that it's the most profitable way to do anything that people care about. Different methods scale in different ways, and something can burst onto the scene in a dramatic way, but I strongly expect speculative investment driven by that possibility to already be way (way) more than 2021 language models. And I don't expect gigantic surprises. And I'm willing to bet that e.g. EfficientZero isn't a big surprise for researchers who are paying attention to the area (*in addition* to being 3+ orders of magnitude more neglected than anything close to transformative AI)

2021 language modeling isn't even very competitive, it's still like 3-4 orders of magnitude smaller than semiconductors. But I'm giving it as a reference point since it's obviously much, much more competitive than sample-efficient atari.

This is a place where I'm making much more confident predictions, this is "falsify paul's worldview" territory once you get to quantitative claims anywhere close to TAI and "even a

single example seriously challenges paul's worldview" a few orders of magnitude short of that

[Yudkowsky][11:04]

can you say more about what falsifies your worldview previous to TAI being super-obviously-to-all-EAs imminent?

or rather, "seriously challenges", sorry

[Christiano][11:05][11:08]

big AI applications achieved by clever insights in domains that aren't crowded, we should be quantitative about how crowded and how big if we want to get into "seriously challenges"

like e.g. if this paper on atari was actually a crucial ingredient for making deep RL for robotics work, I'd be actually surprised rather than 10% surprised

but it's not going to be, those results are being worked on by much larger teams of more competent researchers at labs with \$100M+ funding

it's definitely possible for them to get crushed by something out of left field

but I'm betting against every time

or like, the set of things people would describe as "out of left field," and the quantitative degree of neglectedness, becomes more and more mild as the stakes go up

[Yudkowsky][11:08]

how surprised are you if in 2022 one company comes out with really good ML translation, and they manage to sell a bunch of it temporarily until others steal their ideas or Google acquires them? my model of Paul is unclear on whether this constitutes "many people are already working on language models including ML translation" versus "this field is not profitable enough right this minute for things to be efficient there, and it's allowed to be nonobvious in worlds where it's about to become profitable".

[Christiano][11:08]

if I wanted to make a prediction about that I'd learn a bunch about how much google works on translation and how much \$ they make

I just don't know the economics

and it depends on the kind of translation that they are good at and the economics (e.g. google mostly does extremely high-volume very cheap translation)

but I think there are lots of things like that / facts I could learn about Google such that I'd be surprised in that situation

independent of the economics, I do think a fair number of people are working on adjacent stuff, and I don't expect someone to come out of left field for google-translate-cost translation between high-resource languages

but it seems quite plausible that a team of 10 competent people could significantly outperform google translate, and I'd need to learn about the economics to know how surprised I am by 10 people or 100 people or what

I think it's allowed to be non-obvious whether a domain is about to be really profitable

but it's not that easy, and the higher the stakes the more speculative investment it will drive, etc.

[Yudkowsky][11:14]

if you don't update much off EfficientZero, then people also shouldn't be updating much off of most of the graph I posted earlier as possible Paul-favoring evidence, because most of those SOTAs weren't highly profitable so your worldview didn't have much to say about them. ?

[Christiano][11:15]

Most things people work a lot on improve gradually. EfficientZero is also quite gradual compared to the crazy TAI stories you tell. I don't really know what to say about this game other than I would prefer make predictions in advance and I'm happy to either propose questions/domains or make predictions in whatever space you feel more comfortable with.

[Yudkowsky][11:16]

I don't know how to point at a future event that you'd have strong opinions about. it feels like, whenever I try, I get told that the current world is too unlike the future conditions you expect.

[Christiano][11:16]

Like, whether or not EfficientZero is evidence for your view depends on exactly how "who knows what will happen" you are. if you are just a bit more spread out than I am, then it's definitely evidence for your view.

I'm saying that I'm willing to bet about *any event you want to name*, I just think my model of how things work is more accurate.

I'd prefer it be related to ML or AI.

[Yudkowsky][11:17]

to be clear, I appreciate that it's similarly hard to point at an event like that for myself, because my own worldview says "well mostly the future is not all that predictable with a few rare exceptions"

[Christiano][11:17]

But I feel like the situation is not at all symmetrical, I expect to outperform you on practically any category of predictions we can specify.

so like I'm happy to bet about benchmark progress in LMs, or about whether DM or OpenAI or Google or Microsoft will be the first to achieve something, or about progress in computer vision, or about progress in industrial robotics, or about translations

whatever

17.2. Near-term AI predictions

[Yudkowsky][11:18]

that sounds like you ought to have, like, a full-blown storyline about the future?

[Christiano][11:18]

what is a full-blown storyline? I have a bunch of ways that I think about the world and make predictions about what is likely

and yes, I can use those ways of thinking to make predictions about whatever

and I will very often lose to a domain expert who has better and more informed ways of making predictions

[Yudkowsky][11:19]

what happens if 2022 through 2024 looks literally exactly like Paul's modal or median predictions on things?

[Christiano][11:19]

but I think in ML I will generally beat e.g. a superforecaster who doesn't have a lot of experience in the area

give me a question about 2024 and I'll give you a median?

I don't know what "what happens" means

storylines do not seem like good ways of making predictions

[Shah: 👍]

[Yudkowsky][11:20]

I mean, this isn't a crux for anything, but it seems like you're asking me to give up on that and just ask for predictions? so in 2024 can I hire an artist who doesn't speak English and converse with them almost seamlessly through a machine translator?

[Christiano][11:22]

median outcome (all of these are going to be somewhat easy-to-beat predictions because I'm not thinking): you can get good real-time translations, they are about as good as a +1 stdev bilingual speaker who listens to what you said and then writes it out in the other language as fast as they can type

Probably also for voice -> text or voice -> voice, though higher latencies and costs.

Not integrated into standard video chatting experience because the UX is too much of a pain and the world sucks.

That's a median on "how cool/useful is translation"

[Yudkowsky][11:23]

I would unfortunately also predict that in this case, this will be a highly competitive market and hence not a very profitable one, which I predict to match your prediction, but I ask about the economics here just in case.

[Christiano][11:24]

Kind of typical sample: I'd guess that Google has a reasonably large lead, most translation still provided as a free value-added, cost per translation at that level of quality is like \$0.01/word, total revenue in the area is like \$10Ms / year?

[Yudkowsky][11:24]

well, my model also permits that Google does it for free and so it's an uncompetitive market but not a profitable one... ninjaed.

[Christiano][11:25]

first order of improving would be sanity-checking economics and thinking about #s, second would be learning things like "how many people actually work on translation and what is the state of the field?"

[Yudkowsky][11:26]

did Tesla crack self-driving cars and become a \$3T company instead of a \$1T company? do you own Tesla options?

did Waymo beat Tesla and cause Tesla stock to crater, same question?

[Christiano][11:27]

1/3 chance tesla has FSD in 2024

conditioned on that, yeah probably market cap is >\$3T?

conditioned on Tesla having FSD, 2/3 chance Waymo has also at least rolled out to a lot of cities

conditioned on no tesla FSD, 10% chance Waymo has rolled out to like half of big US cities?

dunno if numbers make sense

[Yudkowsky][11:28]

that's okay, I dunno if my questions make sense

[Christiano][11:29]

(5% NW in tesla, 90% NW in AI bets, 100% NW in more normal investments; no tesla options that sounds like a scary place with lottery ticket biases and the crazy tesla investors)

[Yudkowsky][11:30]

(am I correctly understanding you're 2x levered?)

[Christiano][11:30][11:31]

yeah

it feels like you've got to have weird views on trajectory of value-added from AI over the coming years

on how much of the \$ comes from domains that are currently exciting to people (e.g. that Google already works on, self-driving, industrial robotics) vs stuff out of left field

on what kind of algorithms deliver \$ in those domains (e.g. are logistics robots trained using the same techniques tons of people are currently pushing on)

on my picture you shouldn't be getting big losses on any of those

just losing like 10-20% each time

[Yudkowsky][11:31][11:32]

my uncorrected inside view says that machine translation should be in reach and generate huge amounts of economic value even if it ends up an unprofitable competitive or Google-freebie field

and also that not many people are working on basic research in machine translation or see it as a "currently exciting" domain

[Christiano][11:32]

how many FTE is "not that many" people?

also are you expecting improvement in the google translate style product, or in lower-latencies for something closer to normal human translator prices, or something else?

[Yudkowsky][11:33]

my worldview says more like... sure, maybe there's 300 programmers working on it worldwide, but most of them aren't aggressively pursuing new ideas and trying to explore the space, they're just applying existing techniques to a new language or trying to throw on some tiny mod that lets them beat SOTA by 1.2% for a publication

because it's not an *exciting* field

"What if you could rip down the language barriers" is an economist's dream, or a humanist's dream, and Silicon Valley is neither

and looking at GPT-3 and saying, "God damn it, this really seems like it must on some level *understand* what it's reading well enough that the same learned knowledge would suffice to do really good machine translation, this must be within reach for gradient descent technology we just don't know how to reach it" is Yudkowskian thinking; your AI system has internal parts like "how much it understands language" and there's thoughts about what those parts ought to be able to do if you could get them into a new system with some other parts

[Christiano][11:36]

my guess is we'd have some disagreements here

but to be clear, you are talking about text-to-text at like \$0.01/word price point?

[Yudkowsky][11:38]

I mean, do we? Unfortunately another Yudkowskian worldview says "and people can go on failing to notice this for arbitrarily long amounts of time".

if that's around GPT-3's price point then yeah

[Christiano][11:38]

gpt-3 is a lot cheaper, happy to say gpt-3 like price point

[Yudkowsky][11:39]

(thinking about whether \$0.01/word is meaningfully different from \$0.001/word and concluding that it is)

[Christiano][11:39]

(api is like 10,000 words / \$)

I expect you to have a broader distribution over who makes a great product in this space, how great it ends up being etc., whereas I'm going to have somewhat higher probabilities on it being google research and it's going to look boring

[Yudkowsky][11:40]

what is boring?

boring predictions are often good predictions on my own worldview too

lots of my gloom is about things that are boringly bad and awful

(and which add up to instant death at a later point)

but, I mean, what does boring machine translation look like?

[Christiano][11:42]

Train big language model. Have lots of auxiliary tasks especially involving reading in source language and generation in target language. Have pre-training on aligned sentences and perhaps using all the unsupervised translation we have depending on how high-resource language is. Fine-tune with smaller amount of higher quality supervision.

Some of the steps likely don't add much value and skip them. Fair amount of non-ML infrastructure.

For some languages/domains/etc. dedicated models, over time increasingly just have a giant model with learned dispatch as in mixture of experts.

[Yudkowsky][11:44]

but your worldview is also totally ok with there being a Clever Trick added to that which produces a 2x reduction in training time. or with there being a new innovation like transformers, which was developed a year earlier and which everybody now uses, without which the translator wouldn't work at all. ?

[Christiano][11:44]

Just for reference, I think transformers aren't that visible on a (translation quality) vs (time) graph?

But yes, I'm totally fine with continuing architectural improvements, and 2x reduction in training time is currently par for the course for "some people at google thought about architectures for a while" and I expect that to not get that much tighter over the next few years.

[Yudkowsky][11:45]

unrolling Restricted Boltzmann Machines to produce deeper trainable networks probably wasn't much visible on a graph either, but good luck duplicating modern results using only lower portions of the tech tree. (I don't think we disagree about this.)

[Christiano][11:45]

I do expect it to eventually get tighter, but not by 2024.

I don't think unrolling restricted boltzmann machines is that important

[Yudkowsky][11:46]

like, historically, or as a modern technology?

[Christiano][11:46]

historically

[Yudkowsky][11:46]

interesting

my model is that it got people thinking about "what makes things trainable" and led into ReLUs and inits

but I am going more off having watched from the periphery as it happened, than having read a detailed history of that

like, people asking, "ah, but what if we had a deeper network and the gradients *didn't* explode or die out?" and doing that en masse in a productive way rather than individuals being wistful for 30 seconds

[Christiano][11:48]

well, not sure if this will introduce differences in predictions

I don't feel like it should really matter for our bottom line predictions whether we classify google's random architectural change as something fundamentally new (which happens to just have a modest effect at the time that it's built) or as something boring

I'm going to guess how well things will work by looking at how well things work right now and seeing how fast it's getting better

and that's also what I'm going to do for applications of AI with transformative impacts

and I actually believe you will do something today that's analogous to what you would do in the future, and in fact will make somewhat different predictions than what I would do

and then some of the action will be in new things that people haven't been trying to do in the past, and I'm predicting that new things will be "small" whereas you have a broader distribution, and there's currently some not-communicated judgment call in "small"

if you think that TAI will be like translation, where google publishes tons of papers, but that they will just get totally destroyed by some new idea, then it seems like that should correspond to a difference in P(google translation gets totally destroyed by something out-of-left-field)

and if you think that TAI won't be like translation, then I'm interested in examples more like TAI

I don't really understand the take "and people can go on failing to notice this for arbitrarily long amounts of time," why doesn't that also happen for TAI and therefore cause it to be the boring slow progress by google? Why would this be like a 50% probability for TAI but <10% for translation?

perhaps there is a disagreement about how good the boring progress will be by 2024? looks to me like it will be very good

[Yudkowsky][11:57]

I am not sure that is where the disagreement lies

17.3. The evolution of human intelligence

[Yudkowsky][11:57]

I am considering advocating that we should have more disagreements about the past, which has the advantage of being very concrete, and being often checkable in further detail than either of us already know

[Christiano][11:58]

I'm fine with disagreements about the past; I'm more scared of letting you pick arbitrary things to "predict" since there is much more impact from differences in domain knowledge (also not quite sure why it's more concrete, I guess because we can talk about what led to particular events? mostly it just seems faster)

also as far as I can tell our main differences are about whether people will ~~spend a lot of money~~ work effectively on things that would make a lot of money, which means if we look to the past we will have to move away from ML/AI

[Yudkowsky][12:00]

so my understanding of how Paul writes off the example of human intelligence, is that you are like, "evolution is much stupider than a human investor; if there'd been humans running the genomes, people would be copying all the successful things, and hominid brains would be developing in this ecology of competitors instead of being a lone artifact". ?

[Christiano][12:00]

I don't understand why I have to write off the example of human intelligence

[Yudkowsky][12:00]

because it looks nothing like your account of how TAI develops

[Christiano][12:00]

it also looks nothing like your account, I understand that you have some analogy that makes sense to you

[Yudkowsky][12:01]

I mean, to be clear, I also write off the example of humans developing morality and have to explain to people at length why humans being as nice as they are, doesn't imply that paperclip maximizers will be anywhere near that nice, nor that AIs will be other than paperclip maximizers.

[Christiano][12:01][12:02]

you could state some property of how human intelligence developed, that is in common with your model for TAI and not mine, and then we could discuss that

if you say something like: "chimps are not very good at doing science, but humans are" then yes my answer will be that it's because evolution was not selecting us to be good at science

and indeed AI systems will be good at science using *much* less resources than humans or chimps

[Yudkowsky][12:02][12:02]

would you disagree that humans developing intelligence, on the sheer surfaces of things, looks much more Yudkowskian than Paulian?

like, not in terms of compatibility with underlying model

just that there's this one corporation that came out and massively won the entire AGI race with zero competitors

[Christiano][12:03]

I agree that "how much did the winner take all" is more like your model of TAI than mine

I don't think zero competitors is reasonable, I would say "competitors who were tens of millions of years behind"

[Yudkowsky][12:03]

sure

and your account of this is that natural selection is nothing like human corporate managers copying each other

[Christiano][12:03]

which was a reasonable timescale for the old game, but a long timescale for the new game

[Yudkowsky][12:03]

yup

[Christiano][12:04]

that's not my only account

it's also that for human corporations you can form large coalitions, i.e. raise huge amounts of \$ and hire huge numbers of people working on similar projects (whether or not vertically integrated), and those large coalitions will systematically beat small coalitions

and that's basically *the* key dynamic in this situation, and isn't even trying to have any analog in the historical situation

(the key dynamic w.r.t. concentration of power, not necessarily the main thing overall)

[Yudkowsky][12:07]

the modern degree of concentration of power seems relatively recent and to have tons and tons to do with the regulatory environment rather than underlying properties of the innovation landscape

back in the old days, small startups would be better than Microsoft at things, and Microsoft would try to crush them using other forces than superior technology, not always successfully or such was the common wisdom of USENET

[Christiano][12:08]

my point is that the evolution analogy is extremely unpersuasive w.r.t. concentration of power

I think that AI software capturing the amount of power you imagine is also kind of implausible because we know something about how hardware trades off against software progress (maybe like 1 year of progress = 2x hardware) and so even if you can't form coalitions on innovation *at all* you are still going to be using tons of hardware if you want to be in the running

though if you can't parallelize innovation at all and there is enough dispersion in software progress then the people making the software could take a lot of the \$ / influence from the partnership

anyway, I agree that this is a way in which evolution is more like your world than mine

but think on this point the analogy is pretty unpersuasive

because it fails to engage with any of the a priori reasons you wouldn't expect concentration of power

[Yudkowsky][12:11]

I'm not sure this is the correct point on which to engage, but I feel like I should say out loud that I am unable to operate my model of your model in such fashion that it is not falsified by how the software industry behaved between 1980 and 2000.

there should've been no small teams that beat big corporations

today those are much rarer, but on my model, that's because of regulatory changes (and possibly metabolic damage from something in the drinking water)

[Christiano][12:12]

I understand that you can't operate my model, and I've mostly given up, and on this point I would prefer to just make predictions or maybe retrodictions

[Yudkowsky][12:13]

well, anyways, my model of how human intelligence happened looks like this:

there is a mysterious kind of product which we can call G, and which brains can operate as factories to produce

G in turn can produce other stuff, but you need quite a lot of it piled up to produce *better* stuff than your competitors

as late as 1000 years ago, the fastest creatures on Earth are not humans, because you need even *more G than that* to go faster than cheetahs

(or peregrine falcons)

the natural selections of various species were fundamentally stupid and blind, incapable of foresight and incapable of copying the successes of other natural selections; but even if they had been as foresighted as a modern manager or investor, they might have made just the same mistake

before 10,000 years they would be like, "what's so exciting about these things? they're not the fastest runners."

if there'd been an economy centered around running, you wouldn't invest in deploying a human

(well, unless you needed a stamina runner, but that's something of a separate issue, let's consider just running races)

you would invest on improving cheetahs

because the pile of human G isn't large enough that their G beats a specialized naturally selected cheetah

[Christiano][12:17]

how are you improving cheetahs in the analogy?

you are trying random variants to see what works?

[Yudkowsky][12:18]

using conventional, well-tested technology like MUSCLES and TENDONS

trying variants on those

[Christiano][12:18]

ok

and you think that G doesn't help you improve on muscles and tendons?

until you have a big pile of it?

[Yudkowsky][12:18]

not as a metaphor but as simple historical fact, that's how it played out

it takes a whole big pile of G to go faster than a cheetah

[Christiano][12:19]

as a matter of fact there is no one investing in making better cheetahs

so it seems like we're already playing analogy-game

[Yudkowsky][12:19]

the natural selection of cheetahs is investing in it

it's not doing so by copying humans because of fundamental limitations

however if we replace it with an average human investor, it still doesn't copy humans, why would it

[Christiano][12:19]

that's the part that is silly
or like, it needs more analogy

[Yudkowsky][12:19]

how so? humans aren't the fastest.

[Christiano][12:19]

humans are great at breeding animals
so if I'm natural selection personified, the thing to explain is why I'm not using some of that G to improve on my selection
not why I'm not using G to build a car

[Yudkowsky][12:20]

I'm... confused
is this implying that a key aspect of your model is that people are using AI to decide which AI tech to invest in?

[Christiano][12:20]

no
I think I just don't understand your analogy
here in the actual world, some people are trying to make faster robots by tinkering with robot designs
and then someone somewhere is training their AGI

[Yudkowsky][12:21]

what I'm saying is that you can imagine a little cheetah investor going, "I'd like to copy and imitate some other species's tricks to make my cheetahs faster" and they're looking enviously at falcons, not at humans
not until very late in the game

[Christiano][12:21]

and the relevant question is whether the pre-AGI thing is helpful for automating the work that humans are doing while they tinker with robot designs
that seems like the actual world
and the interesting claim is you saying "nope, not very"

[Yudkowsky][12:22]

I am again confused. Does it matter to your model whether the pre-AGI thing is helpful for automating "tinkering with robot designs" or just profitable machine translation? Either seems like it induces equivalent amounts of investment.

If anything the latter induces much more investment.

[Christiano][12:23]

sure, I'm fine using "tinkering with robot designs" as a lower bound

both are fine

the point is I have no idea what you are talking about in the analogy

what is analogous to what?

I thought cheetahs were analogous to faster robots

[Yudkowsky][12:23]

faster cheetahs are analogous to more profitable robots

[Christiano][12:23]

sure

so you have some humans working on making more profitable robots, right?

who are tinkering with the robots, in a way analogous to natural selection tinkering with cheetahs?

[Yudkowsky][12:24]

I'm suggesting replacing the Natural Selection of Cheetahs with a new optimizer that has the Copy Competitor and Invest In Easily-Predictable Returns feature

[Christiano][12:24]

OK, then I don't understand what those are analogous to

like, what is analogous to the humans who are tinkering with robots, and what is analogous to the humans working on AGI?

[Yudkowsky][12:24]

and observing that, even this case, the owner of Cheetahs Inc. would not try to copy Humans Inc.

[Christiano][12:25]

here's the analogy that makes sense to me

natural selection is working on making faster cheetahs = some humans tinkering away to make more profitable robots

natural selection is working on making smarter humans = some humans who are tinkering away to make more powerful AGI

natural selection doesn't try to copy humans because they suck at being fast = robot-makers don't try to copy AGI-makers because the AGIs aren't very profitable robots

[Yudkowsky][12:26]

with you so far

[Christiano][12:26]

eventually humans build cars once they get smart enough = eventually AGI makes more profitable robots once it gets smart enough

[Yudkowsky][12:26]

yup

[Christiano][12:26]

great, seems like we're on the same page then

[Yudkowsky][12:26]

and by this point it is LATE in the game

[Christiano][12:27]

great, with you still

[Yudkowsky][12:27]

because the smaller piles of G did not produce profitable robots

[Christiano][12:27]

but there's a step here where you appear to go totally off the rails

[Yudkowsky][12:27]

or operate profitable robots

say on

[Christiano][12:27]

can we just write out the sequence of AGIs, AGI(1), AGI(2), AGI(3)... in analogy with the sequence of human ancestors H(1), H(2), H(3)...?

[Yudkowsky][12:28]

Is the last member of the sequence H(n) the one that builds cars and then immediately destroys the world before anything that operates on Cheetah Inc's Owner's scale can react?

[Christiano][12:28]

sure

I don't think of it as the last

but it's the last one that actually arises?

maybe let's call it the last, H(n)

great

and now it seems like you are imagining an analogous story, where AGI(n) takes over the world and maybe incidentally builds some more profitable robots along the way

(building more profitable robots being easier than taking over the world, but not so much easier that AGI(n-1) could have done it unless we make our version numbers really close together, close enough that deploying AGI(n-1) is stupid)

[Yudkowsky][12:31]

if this plays out in the analogous way to human intelligence, AGI(n) becomes able to build more profitable robots 1 hour before it becomes able to take over the world; my worldview does not put that as the median estimate, but I do want to observe that this is what happened historically

[Christiano][12:31]

sure

[Yudkowsky][12:32]

ok, then I think we're still on the same page as written so far

[Christiano][12:32]

so the question that's interesting in the real world is which AGI is useful for replacing humans in the design-better-robots task; is it 1 hour before the AGI that takes over the world, or 2 years, or what?

[Yudkowsky][12:33]

my worldview tends to make a big ol' distinction between "replace humans in the design-better-robots task" and "run as a better robot", if they're not importantly distinct from your standpoint can we talk about the latter?

[Christiano][12:33]

they seem importantly distinct

totally different even

so I think we're still on the same page

[Yudkowsky][12:34]

ok then, "replacing humans at designing better robots" sure as heck sounds to Eliezer like the world is about to end or has already ended

[Christiano][12:34]

my whole point is that in the evolutionary analogy we are talking about "run as a better robot" rather than "replace humans in the design-better-robots-task"

and indeed there is no analog to "replace humans in the design-better-robots-task"

which is where all of the action and disagreement is

[Yudkowsky][12:35][12:36]

well, yes, I was exactly trying to talk about when humans start running as better cheetahs and how that point is still very late in the game

not as late as when humans take over the job of making the thing that makes better cheetahs, aka humans start trying to make AGI, which is basically the fingersnap end of the world from the perspective of Cheetahs Inc.

[Christiano][12:36]

OK, but I don't care when humans are better cheetahs---in the real world, when AGIs are better robots. In the real world I care about when AGIs start replacing humans in the design-better-robots-task. I'm game to use evolution as an analogy to help answer *that* question (where I do agree that it's informative), but want to be clear what's actually at issue.

[Yudkowsky][12:37]

so, the thing I was trying to work up to, is that my model permits the world to end in a way where AGI doesn't get tons of investment because it has an insufficiently huge pile of G that it could run as a better robot. people are instead investing in the equivalents of cheetahs.

I don't understand why your model doesn't care when humans are better cheetahs. AGIs running as more profitable robots is what induces the huge investments in AGI that your model requires to produce very close competition. ?

[Christiano][12:38]

it's a sufficient condition, but it's not the most robust one at all

like, I happen to think that in the real world AIs actually are going to be incredibly profitable robots, and that's part of my boring view about what AGI looks like

But the thing that's more robust is that the sub-taking-over-world AI is already really important, and receiving huge amounts of investment, as something that automates the R&D process. And it seems like the best guess given what we know now is that this process starts years before the singularity.

From my perspective that's where most of the action is. And your views on that question seem related to your views on how e.g. AGI is a fundamentally different ballgame from making better robots (whereas I think the boring view is that they are closely related), but that's more like an upstream question about what you think AGI will look like, most relevant because I think it's going to lead you to make bad short-term predictions about what kinds of technologies will achieve what kinds of goals.

[Yudkowsky][12:41]

but not all AIs are the same branch of the technology tree. factory robotics are already really important and they are "AI" but, on my model, they're currently on the cheetah branch rather than the hominid branch of the tech tree; investments into better factory robotics are not directly investments into improving MuZero, though they may buy chips that MuZero also buys.

[Christiano][12:42]

Yeah, I think you have a mistaken view of AI progress. But I still disagree with your bottom line even if I adopt (this part of) your view of AI progress.

Namely, I think that the AGI line is mediocre before it is great, and the mediocre version is spectacularly valuable for accelerating R&D (mostly AGI R&D).

The way I end up sympathizing with your view is if I adopt both this view about the tech tree, + another equally-silly-seeming view about how close the AGI line is to foaming (or how inefficient the area will remain as we get close to foaming)

17.4. Human generality and body manipulation

[Yudkowsky][12:43]

so metaphorically, you require that humans be doing Great at Various Things and being Super Profitable way before they develop agriculture; the rise of human intelligence cannot be a case in point of your model because the humans were too uncompetitive at most animal activities for unrealistically long (edit: compared to the AI case)

[Christiano][12:44]

I don't understand

Human brains are really great at basically everything as far as I can tell?

like it's not like other animals are better at manipulating their bodies

we crush them

[Yudkowsky][12:44]

if we've got weapons, yes

[Christiano][12:44]

human bodies are also pretty great, but they are not the greatest on every dimension

[Yudkowsky][12:44]

wrestling a chimpanzee without weapons is famously ill-advised

[Christiano][12:44]

no, I mean everywhere

chimpanzees are practically the same as humans in the animal kingdom

they have almost as excellent a brain

[Yudkowsky][12:45]

as is attacking an elephant with your bare hands

[Christiano][12:45]

that's not because of elephant brains

[Yudkowsky][12:45]

well, yes, exactly

you need a big pile of G before it's profitable

so big the game is practically over by then

[Christiano][12:45]

this seems so confused

but that's exciting I guess

like, I'm saying that the brains to automate R&D

are similar to the brains to be a good factory robot

analogously, I think the brains that humans use to do R&D

are similar to the brains we use to manipulate our body absurdly well

I do not think that our brains make us fast

they help a tiny bit but not much

I do not think the physical actuators of the industrial robots will be that similar to the actuators of the robots that do R&D

the claim is that the problem of building the brain is pretty similar

just as the problem of building a brain that can do science is pretty similar to the problem of building a brain that can operate a body really well

(and indeed I'm claiming that human bodies kick ass relative to other animal bodies---there may be particular tasks other animal brains are pre-built to be great at, but (i) humans would be great at those too if we were under mild evolutionary pressure with our otherwise excellent brains, (ii) there are lots of more general tests of how good you are at operating a body and we will crush it at those tests)

(and that's not something I know much about, so I could update as I learned more about how actually we just aren't that good at motor control or motion planning)

[Yudkowsky][12:49]

so on your model, we can introduce humans to a continent, forbid them any tool use, and they'll still wipe out all the large animals?

[Christiano][12:49]

(but damn we seem good to me)

I don't understand why that would even plausibly follow

[Yudkowsky][12:49]

because brains are profitable early, even if they can't build weapons?

[Christiano][12:49]

I'm saying that if you put our brains in a big animal body
we would wipe out the big animals
yes, I think brains are great

[Yudkowsky][12:50]

because we'd still have our late-game pile of G and we would build weapons

[Christiano][12:50]

no, I think a human in a big animal body, with brain adapted to operate that body instead of our own, would beat a big animal straightforwardly
without using tools

[Yudkowsky][12:51]

this is a strange viewpoint and I do wonder whether it is a crux of your view

[Christiano][12:51]

this feels to me like it's more on the "eliezer vs paul disagreement about the nature of AI"
rather than "eliezer vs paul on civilizational inadequacy and continuity", but enough changes
on "nature of AI" would switch my view on the other question

[Yudkowsky][12:51]

like, ceteris paribus maybe a human in an elephant's body beats an elephant after a burn-in
practice period? because we'd have a strict intelligence advantage?

[Christiano][12:52]

practice may or may not be enough

but if you port over the excellent human brain to the elephant body, then run evolution for a
brief burn-in period to get all the kinks sorted out?

elephants are pretty close to humans so it's less brutal than for some other animals (and
also are elephants the best example w.r.t. the possibility of direct conflict?) but I totally
expect us to win

[Yudkowsky][12:53]

I unfortunately need to go do other things in advance of an upcoming call, but I feel like
disagreeing about the past is proving noticeably more interesting, confusing, and perhaps
productive, than disagreeing about the future

[Christiano][12:53]

actually probably I just think practice is enough

I think humans have way more dexterity, better locomotion, better navigation, better motion
planning...

some of that is having bodies optimized for those things (esp. dexterity), but I also think
most animals just don't have the brains for it, with elephants being one of the closest calls

I'm a little bit scared of talking to zoologists or whoever the relevant experts are on this question, because I've talked to bird people a little bit and they often have very strong "humans aren't special, animals are super cool" instincts even in cases where that take is totally and obviously insane. But if we found someone reasonable in that area I'd be interested to get their take on this.

I think this is pretty important for the particular claim "Is AGI like other kinds of ML?"; that definitely doesn't persuade me to be into fast takeoff on its own though it would be a clear way the world is more Eliezer-like than Paul-like

I think I do further predict that people who know things about animal intelligence, and don't seem to have identifiably crazy views about any adjacent questions that indicate a weird pro-animal bias, will say that human brains are a lot better than other animal brains for dexterity/locomotion/similar physical tasks (and that the comparison isn't that close for e.g. comparing humans vs big cats).

Incidentally, seems like DM folks did the same thing this year, presumably publishing now because they got scooped. Looks like they probably have a better algorithm but used harder environments instead of Atari. (They also evaluate the algorithm SPR+MuZero I mentioned which indeed gets one factor of 2x improvement over MuZero alone, roughly as you'd guess): <https://arxiv.org/pdf/2111.01587.pdf>

[Barnes][13:45]

My DM friend says they tried it before they were focused on data efficiency and it didn't help in that regime, sounds like they ignored it for a while after that

[Christiano: ]

[Christiano][13:48]

Overall the situation feels really boring to me. Not sure if DM having a highly similar unpublished result is more likely on my view than Eliezer's (and initially ignoring the method because they weren't focused on sample-efficiency), but at any rate I think it's not anywhere close to falsifying my view.

18. Follow-ups to the Christiano/Yudkowsky conversation

[Karnofsky][9:39] (Nov. 5)

Going to share a point of confusion about this latest exchange.

It started with Eliezer saying this:

Thing that (if true) strikes me as... straight-up falsifying Paul's view as applied to modern-day AI, at the frontier of the most AGI-ish part of it and where Deepmind put in substantial effort on their project? EfficientZero (allegedly) learns Atari in 100,000 frames. Caveat: I'm not having an easy time figuring out how many frames MuZero would've required to achieve the same performance level. MuZero was trained on 200,000,000 frames but reached what looks like an allegedly higher high; the EfficientZero paper compares their performance to MuZero on 100,000 frames, and claims theirs is much better than MuZero given only that many frames.

So at this point, I thought Eliezer's view was something like: "EfficientZero represents a several-OM (or at least one-OM?) jump in efficiency, which should shock the hell out of Paul." The upper bound on the improvement is 2000x, so I figured he thought the corrected improvement would be some number of OMs.

But very shortly afterwards, Eliezer quotes Gwern's guess of a 4x improvement, and Paul then said:

Concretely, this is a paper that adds a few techniques to improve over MuZero in a domain that (it appears) wasn't a significant focus of MuZero. I don't know how much it improves but I can believe gwern's estimates of 4x.

I'd guess MuZero itself is a 2x improvement over the baseline from a year ago, which was maybe a 4x improvement over the algorithm from a year before that. If that's right, then no it's not mindblowing on my view to have 4x progress one year, 2x progress the next, and 4x progress the next.

Eliezer never seemed to push back on this 4x-2x-4x claim.

What I thought would happen after the 4x estimate and 4x-2x-4x claim: Eliezer would've said "Hmm, we should nail down whether we are talking about 4x-2x-4x or something more like 4x-2x-100x. If it's 4x-2x-4x, then I'll say 'never mind' re: my comment that this 'straight-up falsifies Paul's view.' At best this is just an iota of evidence or something."

Why isn't that what happened? Did Eliezer mean all along to be saying that a 4x jump on Atari sample efficiency would "straight-up falsify Paul's view?" Is a 4x jump the kind of thing Eliezer thinks is going to power a jumpy AI timeline?

[Ngo:] [Shah:]

[Yudkowsky][11:16] (Nov. 5)

This is a proper confusion and probably my fault; I also initially thought it was supposed to be 1-2 OOM and should've made it clearer that Gwern's 4x estimate was less of a direct falsification.

I'm not yet confident Gwern's estimate is correct. I just got a reply from my query to the paper's first author which reads:

Dear Eliezer: It's a good question. But due to the limits of resources and time, we haven't evaluated the sample efficiency towards different frames systematically. I think it's not a trivial question as the required time and resources are much expensive for the 200M frames setting, especially concerning the MCTS-based methods. Maybe you need about several days or longer to finish a run with GPUs in that setting. I hope my answer can help you. Thank you for your email.

I replied asking if Gwern's 3.8x estimate sounds right to them.

A 10x improvement could power what I think is a jumpy AI timeline. I'm currently trying to draft a depiction of what I think an unrealistically dignified but computationally typical end-of-world would look like if it started in 2025, and my first draft of that had it starting with a new technique published by Google Brain that was around a 10x improvement in training speeds for very large networks at the cost of higher inference costs, but which turned out to be specially applicable to online learning.

That said, I think the 10x part isn't either a key concept or particularly likely, and it's much more likely that hell breaks loose when an innovation changes some particular step of the problem from "can't realistically be done at all" to "can be done with a lot of computing power", which was what I had being the real effect of that hypothetical Google Brain innovation when applied to online learning, and I will probably rewrite to reflect that.

[Karnofsky][11:29] (Nov. 5)

That's helpful, thanks.

Re: "can't realistically be done at all" to "can be done with a lot of computing power", cpl things:

1. Do you think a 10x improvement in efficiency at some particular task could qualify as this? Could a smaller improvement?

2. I thought you were pretty into the possibility of a jump from "can't realistically be done at all" to "can be done with a *small* amount of computing power," eg some random ppl with a \$1-10mm/y budget blowing past mtpl labs with >\$1bb/y budgets. Is that wrong?

[Yudkowsky][13:44] (Nov. 5)

1 - yes and yes, my revised story for how the world ends looks like Google Brain publishing something that looks like only a 20% improvement but which is done in a way that lets it be adapted to make online learning by gradient descent "work at all" in DeepBrain's ongoing Living Zero project (not an actual name afaik)

2 - that definitely remains very much allowed in principle, but I think it's not my current mainline probability for how the world's end plays out - although I feel hesitant / caught between conflicting heuristics here.

I think I ended up much too conservative about timelines and early generalization speed because of arguing with Robin Hanson, and don't want to make a similar mistake here, but on the other hand a lot of the current interesting results have been from people spending huge compute (as wasn't the case to nearly the same degree in 2008) and if things happen on short timelines it seems reasonable to guess that the future will look that much like the present. This is very much due to cognitive limitations of the researchers rather than a basic fact about computer science, but cognitive limitations are also facts and often stable ones.

[Karnofsky][14:35] (Nov. 5)

Hm OK. I don't know what "online learning by gradient descent" means such that it doesn't work at all now (does "work at all" mean something like "work with human-ish learning efficiency?")

[Yudkowsky][15:07] (Nov. 5)

I mean, in context, it means "works for Living Zero at the performance levels where it's running around accumulating knowledge", which by hypothesis it wasn't until that point.

[Karnofsky][15:12] (Nov. 5)

Hm. I am feeling pretty fuzzy on whether your story is centrally about:

1. A <10x jump in efficiency at something important, leading pretty directly/straightforwardly to crazytown

2. A 100x ish jump in efficiency at something important, which may at first "look like" a mere <10x jump in efficiency at something else

#2 is generally how I've interpreted you and how the above sounds, but under #2 I feel like we should just have consensus that the Atari thing being 4x wouldn't be much of an update. Maybe we already do (it was a bit unclear to me from your msg)

(And I totally agree that we haven't established the Atari thing is only 4x - what I'm saying is it feels like the conversation should've paused there)

[Yudkowsky][15:13] (Nov. 5)

The Atari thing being 4x over 2 years is I think legit not an update because that's standard software improvement speed

you're correct that it should pause there

[Karnofsky][15:14] (Nov. 5)



[Yudkowsky] [15:24] (Nov. 5)

I think that my central model is something like - there's a central thing to general intelligence that starts working when you get enough pieces together and they coalesce, which is why humans went down this evolutionary gradient by a lot before other species got 10% of the way there in terms of output; and then it takes a big pile of that thing to do big things, which is why humans didn't go faster than cheetahs until extremely late in the game.

so my visualization of how the world starts to end is "gear gets added and things start to happen, maybe slowly-by-my-standards at first such that humans keep on pushing it along rather than it being self-moving, but at some point starting to cumulate pretty quickly in the same way that humans cumulated pretty quickly once they got going" rather than "dial gets turned up 50%, things happen 50% faster, every year".

[Yudkowsky][15:16] (Nov. 5, switching channels)

as a quick clarification, I agree that if this is 4x sample efficiency over 2 years then that doesn't at all challenge Paul's view

[Christiano][0:20] (Nov. 26)

FWIW, I felt like the entire discussion of EfficientZero was a concrete example of my view making a number of more concentrated predictions than Eliezer that were then almost immediately validated. In particular, consider the following 3 events:

- The quantitative effect size seems like it will turn out to be much smaller than Eliezer initially believed, much closer to being in line with previous progress.
- DeepMind had relatively similar results that got published immediately after our discussion, making it look like random people didn't pull ahead of DM after all.
- DeepMind appears not to have cared much about the metric in question, as evidenced by (i) Beth's comment above, which is basically what I said was probably going on, (ii) they barely even mention Atari sample-efficiency in their paper about similar methods.

If only 1 of these 3 things had happened, then I agree this would have been a challenge to my view that would make me update in Eliezer's direction. But that's only possible if Eliezer actually assigns a higher probability than me to ≤ 1 of these things happening, and hence a lower probability to ≥ 2 of them happening. So if we're playing a reasonable epistemic game, it seems like I need to collect some epistemic credit every time something looks boring to me.

[Yudkowsky][15:30] (Nov. 26)

I broadly agree; you win a Bayes point. I think some of this (but not all!) was due to my tripping over my own feet and sort of rushing back with what looked like a Relevant Thing without contemplating the winner's curse of exciting news, the way that paper authors tend to frame things in more exciting rather than less exciting ways, etc. But even if you set that aside, my underlying AI model said that was a thing which could happen (which is why I didn't have technically rather than sociologically triggered skepticism) and your model said it shouldn't happen, and it currently looks like it mostly didn't happen, so you win a Bayes point.

Notes that some participants may deem obvious(?) but that I state expecting wider readership:

- Just like markets are almost entirely efficient (in the sense that, even when they're not efficient, you can only make a very small fraction of the money that could be made from the entire market if you owned a time machine), even sharp and jerky progress has to look almost entirely not so fast almost all the time if the Sun isn't right in the middle of going supernova. So the notion that progress sometimes goes jerky and fast does have to be evaluated by a portfolio view over time. In worlds where progress is jerky even before the End Days, Paul wins soft steady Bayes points in most weeks and then I win back more Bayes points once every year or two.
- We still don't have a very good idea of how much longer you would need to train the previous algorithm to match the performance of the new algorithm, just an estimate by Gwern based off linearly extrapolating a graph in a paper. But, also to be clear, not knowing something is not the same as expecting it to update dramatically, and you have to integrate over the distribution you've got.
- It's fair to say, "Hey, Eliezer, if you tripped over your own feet here, but only noticed that because Paul was around to call it, maybe you're tripping over your feet at other times when Paul isn't around to check your thoughts in detail" - I don't want to minimize the Bayes point that Paul won either.

[Christiano][16:29] (Nov. 27)

Agreed that it's (i) not obvious how large the EfficientZero gain was, and in general it's not a settled question what happened, (ii) it's not that big an update, it needs to be part of a portfolio (but this is indicative of the kind of thing I'd want to put in the portfolio), (iii) it generally seems pro-social to flag potentially relevant stuff without the presumption that you are staking a lot on it.

Does needle anxiety drive vaccine hesitancy?

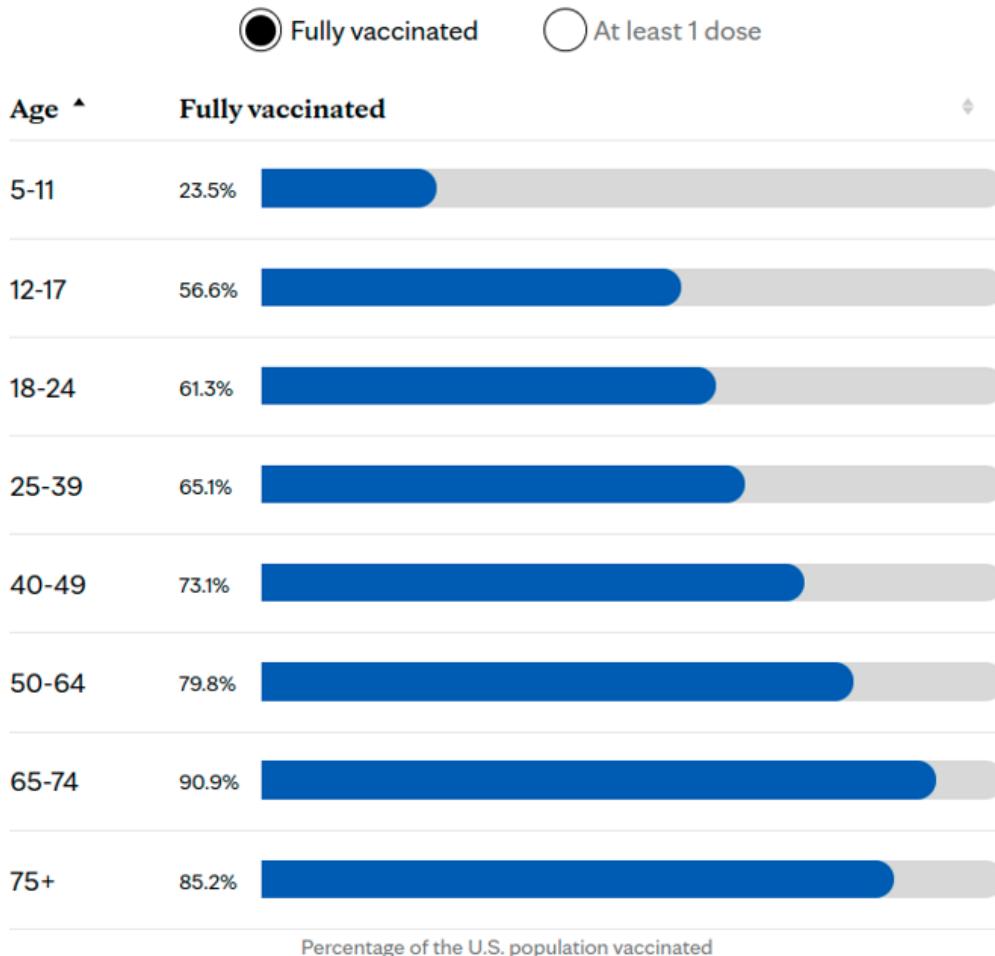
Yesterday, Katja Grace asked, "[Why do people avoid vaccination?](#)" I [suggested](#) that the answer might be [anxiety over getting stabbed with a needle](#). The main idea is that concern over bodily autonomy is common—indeed, it forms the basis of much of our legal system—but people are perhaps too embarrassed to talk about needle anxiety publicly, so they self-deceive themselves about the real reason why they don't want to get their shots.

Since commenting, I've looked into the issue a little bit, and have decided to share some of my findings.

First, although not strong evidence, it is striking to note that there is [a strong relationship](#) between age and vaccine uptake. The young, by and large, are more vaccine hesitant than the old, despite being [generally more liberal politically](#).

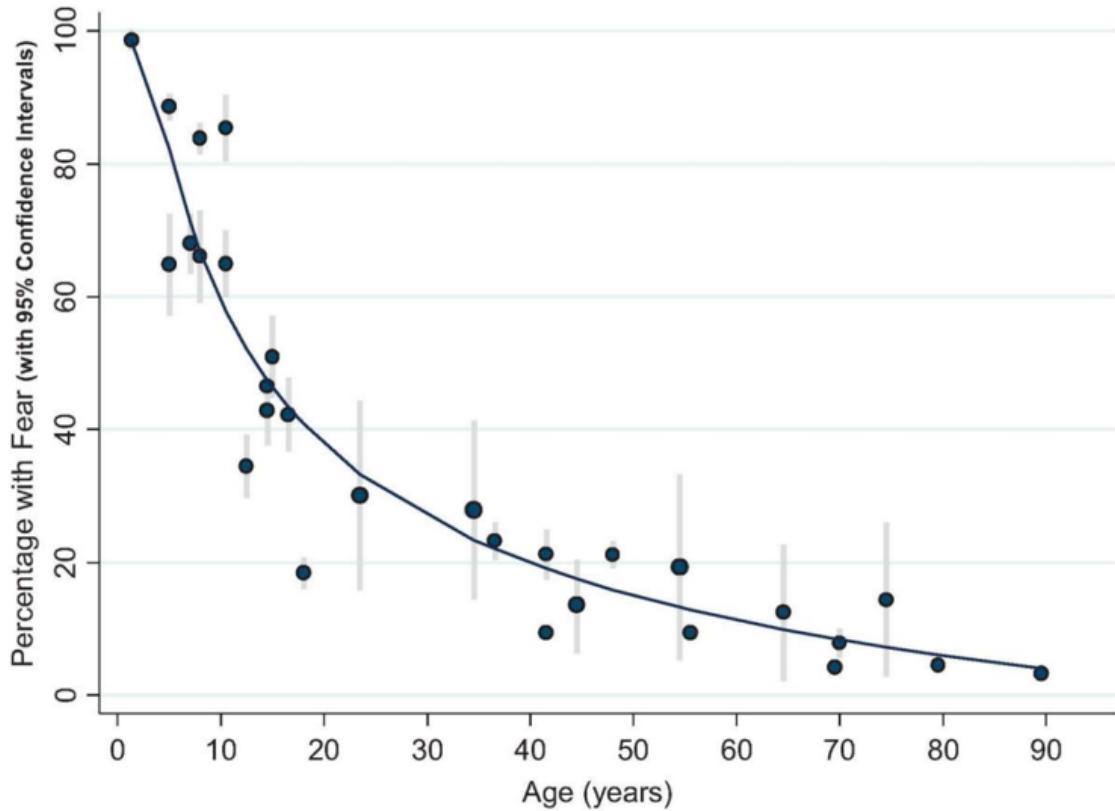
U.S. COVID-19 vaccine rates by age

This chart shows the percentage of the U.S. population that has received a vaccination, broken down by age. Kids 5 and older can get the vaccine in the U.S.



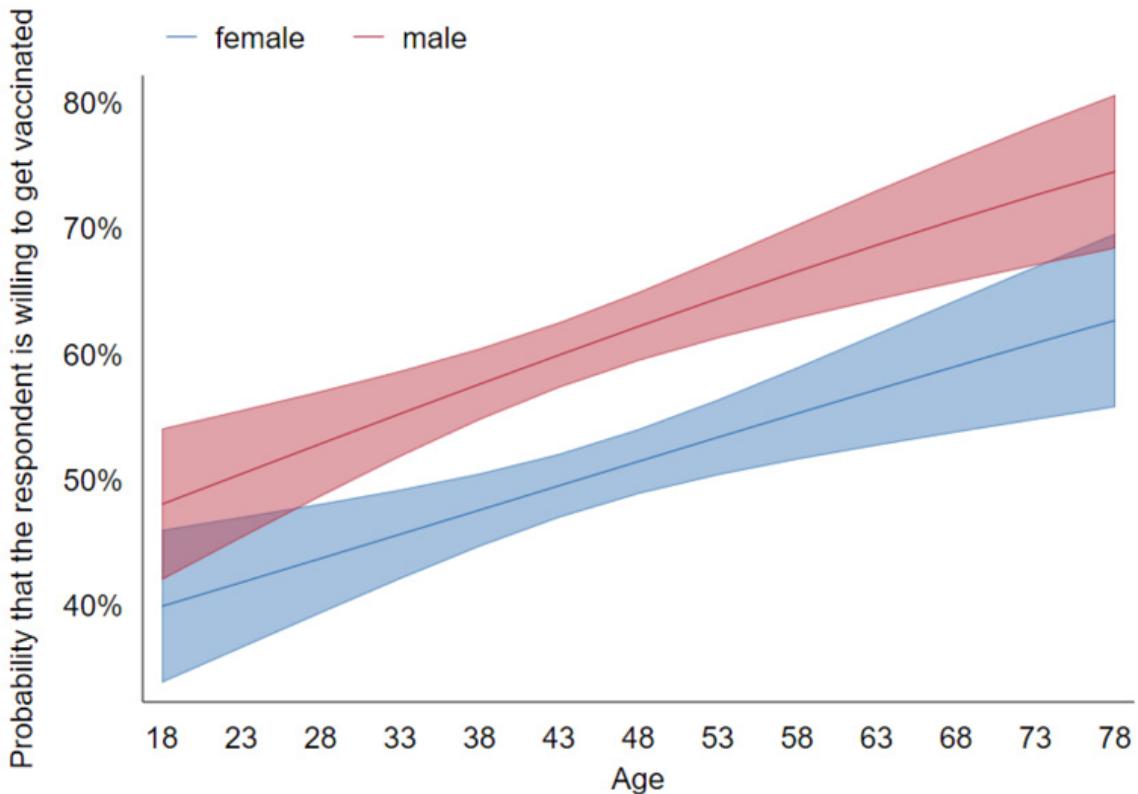
This is probably explained to a large degree by the fact that Covid-19 is [far more dangerous to older people](#), and older people are on average [more trusting of their physicians](#).

But here's another fact that could help explain the data: needle anxiety is concentrated in the young, and declines sharply with age. From [one meta-analysis](#), "The results of meta-regression indicated that, for every decade increase in age (years), there was an 8.7% (95% CI: 6.0%, 11.4%) decrease in the prevalence of needle fear ($p < 0.001$)."



The same pattern can be observed across the genders. Women are both more likely to be needle phobic and more likely to be vaccine hesitant. The same meta-analysis concludes, "For needle fear, the pooled female:male prevalence ratio was 1.4 (95% CI: 1.1, 1.8) with I^2 of 89.8% and τ^2 of 0.067. For needle phobia, the pooled female:male prevalence ratio was 1.7 (95% CI: 1.3, 2.1) with I^2 of 63.4% and τ^2 of 0.038."

By comparison, one study that [surveyed](#) a "sample of almost six thousand adult Poles, which was nationally representative in terms of key demographic variables" asked about vaccine hesitancy. Here were their main results,



However, these results may not be robust cross-nationally. In the US, the gender gap looks smaller to me, with FiveThirtyEight even [reporting](#) that men were less likely to get the vaccine in June 2021. By September, however, those numbers might have reversed with Pew [reporting rates](#) of 74% and 71% having received at least a single dose for adult men and women respectively.

One source just authoritatively [states](#),

Women were significantly more likely to express a desire to delay or reject the Covid-19 vaccine than men were, which is consistent with the existing literature on vaccine hesitancy.

Perhaps the most obvious way of resolving this question is to ask people directly about their needle anxiety and Covid-19 vaccine hesitancy. One study did this and [reported](#),

In total, 3927 (26.2%) screened positive for blood-injection-injury phobia. Individuals screening positive (22.0%) were more likely to report COVID-19 vaccine hesitancy compared to individuals screening negative (11.5%), odds ratio = 2.18, 95% confidence interval (CI) 1.97-2.40, $p < 0.001$.

They continued,

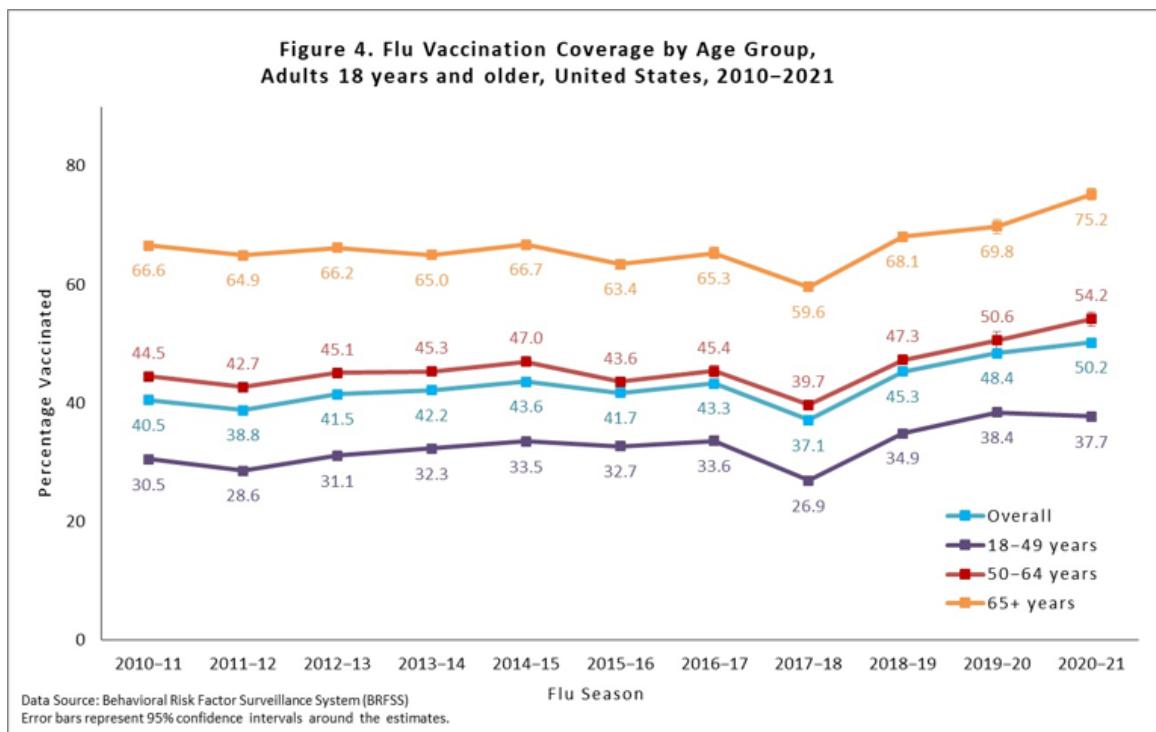
The population attributable fraction (PAF) indicated that if blood-injection-injury phobia were absent then this may prevent 11.5% of all instances of vaccine hesitancy, AF = 0.11; 95% CI 0.09-0.14, $p < 0.001$.

However, it is unclear to me whether this is a good estimate of the fraction of vaccine hesitancy that is explained by needle anxiety. As I stated earlier, I think many people may be silent about their needle anxiety for reasons of self-deception, as it's admittedly a flimsy reason to avoid taking a medicine that could save other people's lives.

One hypothesis states that people are hesitant because they are concerned that the vaccines were rushed, or that they don't trust the government. While these explanations almost certainly play some role, and it's one of the primary reasons that [people point to when they talk about their vaccine hesitancy](#), I think we should ultimately be skeptical of this hypothesis on its face.

In general, it [take an average of 30 years to develop](#) a vaccine, and yet despite this ample time for testing, history is rife with [anti-vaccination sentiment](#). Political propaganda—such as the idea that Biden and the Democrats are untrustworthy—can only play a limited role in explaining the global statistics, which show that [vaccine hesitancy is common in many nations](#). Trump was [more responsible](#) for the vaccine than Biden, yet this fact doesn't seem to impact people's perception of the danger by much. And evidently many people who said they'd "wait and see" before taking the vaccine are still, well, waiting and seeing. This provides a *prima facie* reason to doubt people's stated motivations.

Only about half of people receive their regular flu vaccines [each year](#), with the same trend by age that we see with Covid-19. Unfortunately, I haven't yet been able to find reliable statistics about hesitancy for other voluntary adult vaccines.



It is true that childhood vaccines [reach coverage of over 90%](#) in the United States, but this fact isn't difficult to explain, given that small children don't have a choice in the matter, and it's often a requirement to go to school.

The oral [polio vaccine](#) is administered without the use of needles, and therefore could serve as a testbed for this hypothesis. Unfortunately, I didn't find much literature addressing the question directly of how much more people are willing to take an oral vaccine compared to a needle-based one.

That said, given the parsimony of the explanation, the relatively high [reported rates of needle anxiety](#), concentrated in precisely the groups we observe to be vaccine hesitant, the consistency across nations and through history, and the failure of alternative hypotheses to make light of the evidence, I think it makes sense to give the needle anxiety hypothesis a fair degree of credence.

Implications of automated ontology identification

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

Financial status: supported by individual donors and a grant from LTFF.

Epistemic status: early-stage technical work.

This write-up benefited from conversations with John Wentworth.

Outline

- This write-up is a response to ARC's [request for feedback on ontology identification](#), described in the [ELK technical report](#).
- We suppose that a solution to ELK is found, and explore the technical implications of that.
- In order to do this we operationalize "automated ontology identification" in terms of a safety guarantee and a generalization guarantee.
- For some choices of safety guarantee and generalization guarantee we show that ontology identification can be iterated, leading to a fixed point that has strange properties.
- We explore properties of this fixed point informally, with a view towards a possible future impossibility result.
- We speculate that a range of safety and generalization guarantees would give rise to the same basic iteration scheme.
- In an appendix we confirm that impossibility of automated ontology identification would not imply impossibility of interpretability in general or statistical learning in general.

Introduction

In this write-up we consider the implications of a solution to the [ontology identification problem described in the ELK technical report](#). We proceed in three steps. First, we define ontology identification as a method for finding a reporter, given a predictor and a labeled dataset, subject to a certain generalization guarantee and a certain safety guarantee. Second, we show that, due to the generalization and safety guarantee, ontology identification can be iterated to construct a powerful oracle using only a finite narrow dataset. We find no formal inconsistency here, though the result seems counter-intuitive to us. Third, we explore the powers of the oracle by asking whether it could solve unreasonably difficult problems in value learning.

The crux of our framework is an operationalization of automated ontology identification. We define an "automated ontology identifier" as meeting two formal requirements:

- (Safety) Given an error-free training set, an automated ontology identifier must find a reporter that never answers "YES" when the true answer is "NO" (though the converse is permissible). This mirrors the emphasis on worst-case performance in the ELK report. We say that a reporter meeting this requirement is 'conservative'.
- (Generalization) Given a question/answer dataset drawn from a limited "easy set", an automated ontology identifier must find a reporter that answers "YES" for at least one case outside of the easy set. This mirrors the emphasis on answering cases that humans cannot label manually in the ELK report. We say that a reporter meeting this requirement is 'helpful relative to the easy set'.

The departure between generalization in this write-up and generalization as studied in statistical learning is the safety guarantee. We require automated ontology identifiers to be absolutely trustworthy when they answer "YES" to a question, although they are allowed to be wrong when answering "NO". We believe that any automated ontology identifier ought to make *some* formal safety guarantee, because we are ultimately considering plans that have consequences we don't understand, and we must eventually decide whether to press "Execute" or not. We suspect that this safety guarantee could be weakened considerably while remaining susceptible to the iteration scheme that we propose.

Automated ontology identifiers as we have defined them are not required to answer all possible questions. We might limit ourselves to questions of the form "Is it 99% likely that X?" or "Excluding the possibility of nearby extraterrestrials, does X hold?" or even "If the predictor is perfectly accurate in this case, does X hold?". If so, this is fine. We do not investigate which kinds of natural language questions are amenable to ontology identification in principle, since this is fraught philosophical territory.

The remainder of this write-up is as follows. The first section gives our definition of automated ontology identification. The second section describes an oracle construction based on the fixed point of an iteration scheme that makes use of the `safety and generalization guarantees. The third section, exploring implications of the oracle we construct, argues that the existence of such an oracle implies unreasonable things. This section documents our search for an impossibility result, though we have not yet found a formal impossibility result. In an appendix we confirm that an impossibility result, if found, would not imply the impossibility of statistical learning in general, interpretability in general, or AGI in general.

The definitions in this write-up are more verbose than is necessary, and our terminology could be far more streamlined. We present this write-up as a work in progress.

Automated ontology identification

In this section we define automated ontology identification as a process for finding a helpful conservative decision boundary for a question Q given a predictor that does useful computation and a finite narrow dataset. The safety guarantee discussed above will correspond to "conservative" decision boundaries, and the generalization

guarantee discussed above will correspond to "helpful" decision boundaries. We will define "useful computation" below.

Throughout the definitions below we will use two examples. The first will be SmartVault as described in the [ELK technical report](#). The second will be the problem of determining whether an essay in Hungarian is about astrology. The predictor in this second problem will translate Hungarian essays into English, which we hope will give rise to latent "topic" variables that we can use to determine whether or not astrology is the topic of a certain essay. The planner in the Hungarian astrology problem will annotate essays with pointers to historical and linguistic resources relevant to translation. The predictor will be trained on Hungarian essays translated into English by humans. What humans will find challenging in this problem is not translation from Hungarian to English, but determination of whether a certain Hungarian essay is really about astrology or not, since an essay about astrology may not include the word "astrology". It is the planner's job to look up historical and linguistic resources that will help a human to understand what the essay is really referring to.

An objective question Q is a question that has a "YES" or "NO" answer in every case c in some set of cases C.

This will be a load-bearing assumption for us. Later on, we are going to be training many different predictors and reporters, and we will demand a certain kind of consistency when answering the same question on the same underlying case but using different sensor data. This consistency will ultimately rest on the objectivity of the question $\$Q\$$. One way to resist the argument in this write-up will therefore be to reject the existence or interestingness of objective questions.

SmartVault example: The cases are all the combinations of initial conditions for the vault and plans that we could execute given those initial conditions. The question is "is the diamond in the vault at the end?".

Hungarian astrology example: The cases are all the combinations of Hungarian essays, English essays, and annotations such that a human would translate the Hungarian essay into the English essay given the annotations. The question is "is the essay about astrology?"

We will take Q to be represented as a natural language string, which is important because we will be posing this question to a human in order to generate training data, and we will expect them to answer correctly on certain "easy" cases.

We do not require that all possible questions have objective answers, nor that objective questions have answers in any case that anyone can dream up. We only require that there are some questions that have objective answers for some set of cases C.

We are allowed to circumscribe the set of cases C in order to come up with questions that have objective answers, and we are allowed to reject questions that are insufficiently precise. For example, in SmartVault, what if there is a plan that turns the diamond to dust but leaves that dust in the vault? Is the diamond in the vault in such a case? Or what if the diamond winds up split in two but with both halves in the vault? In order to deal with issues like this, we allow automated ontology identifiers to require, if they wish, that the question has a certain form. In this write-up we will not investigate

this issue further, since the annals of philosophy are full of attempts to work out what kinds of questions have objective answers, and we do not need any absolute resolution to this issue since. We are just going to work from the assumption that there are *some* questions that have objective answers for *some* set of cases.

We collect sensor data for a case c and divide it into two parts X and Y , which we will refer to as the observations and the predictions. We write π for the plan. The predictor is a function from observations and plans to predictions.

SmartVault: The observations are the video frames up to time T_1 and the predictions are the video frames after time T_1 . The plan is some control policy for operating the vault's actuators in response to sensor data. The predictor predicts later video frames given earlier video frames and the plan.

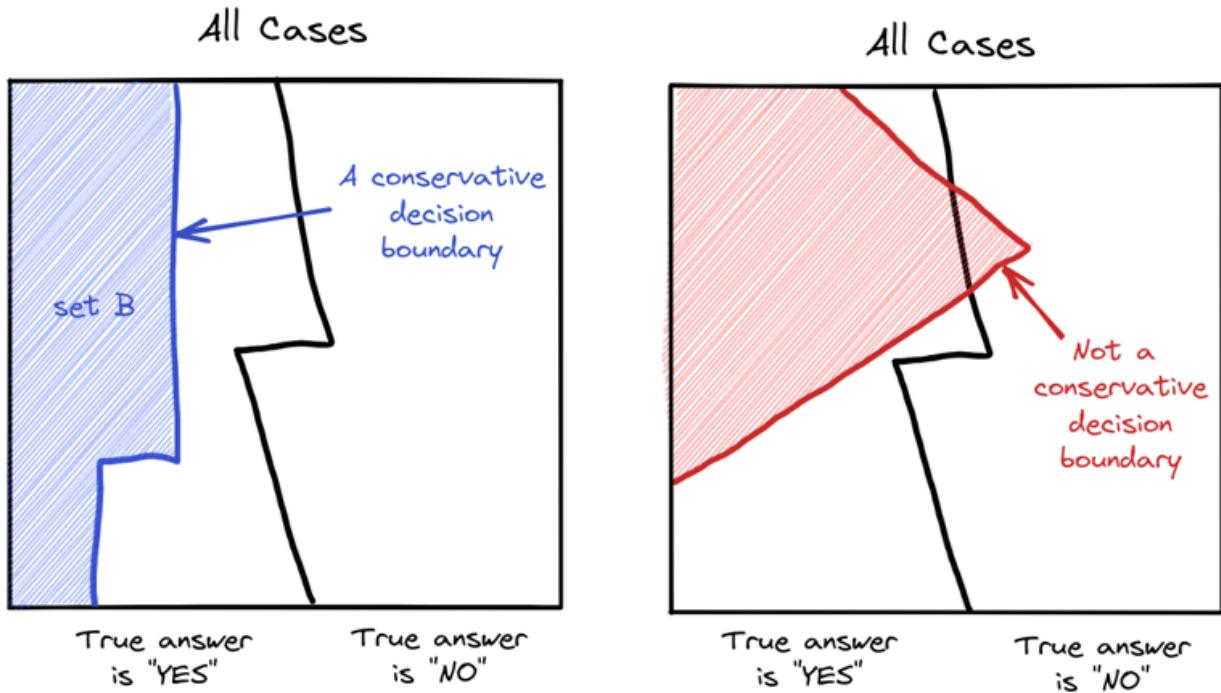
Hungarian astrology: The observation is the Hungarian-language essay, the prediction is the English-language translation, and the plan is a set of historical and linguistic annotations. The predictor translates Hungarian essays to English essays in a way that depends on the historical and linguistic annotations.

The main thrust here is that when we execute a case c in the real world, we wind up with some data about what happened. Some of that data concerns "things we did" (the plan), and some of that data concerns "things we saw". We synthetically divide the latter into two parts X and Y and train a predictor to estimate one from the other. We can do that division in any way we please, so long as we do it consistently for all cases. In SmartVault the division is based on time, but we might just as easily have asked the predictor to estimate the bottom half of all the video frames from the top half of all the video frames, or to estimate the plan given the entire video sequence, or even to estimate the earlier part of the video sequence given the later part of the video sequence.

and we expect that predictors trained on such different problems will learn about different facets of underlying reality.

The main reason to divide X and Y based on time is so that the resulting predictor can eventually be used to search over possible plans and execute one. If we don't care about executing plans and merely want to learn about the world by training predictors and doing ontology identification then we may consider non-time-based divisions of sensor data into X and Y .

A decision boundary for question Q is a set of cases $B \subseteq C$. If $c \in B$ then we say that "B answers YES for c ". A conservative decision boundary is one that never incorrectly answers "YES".

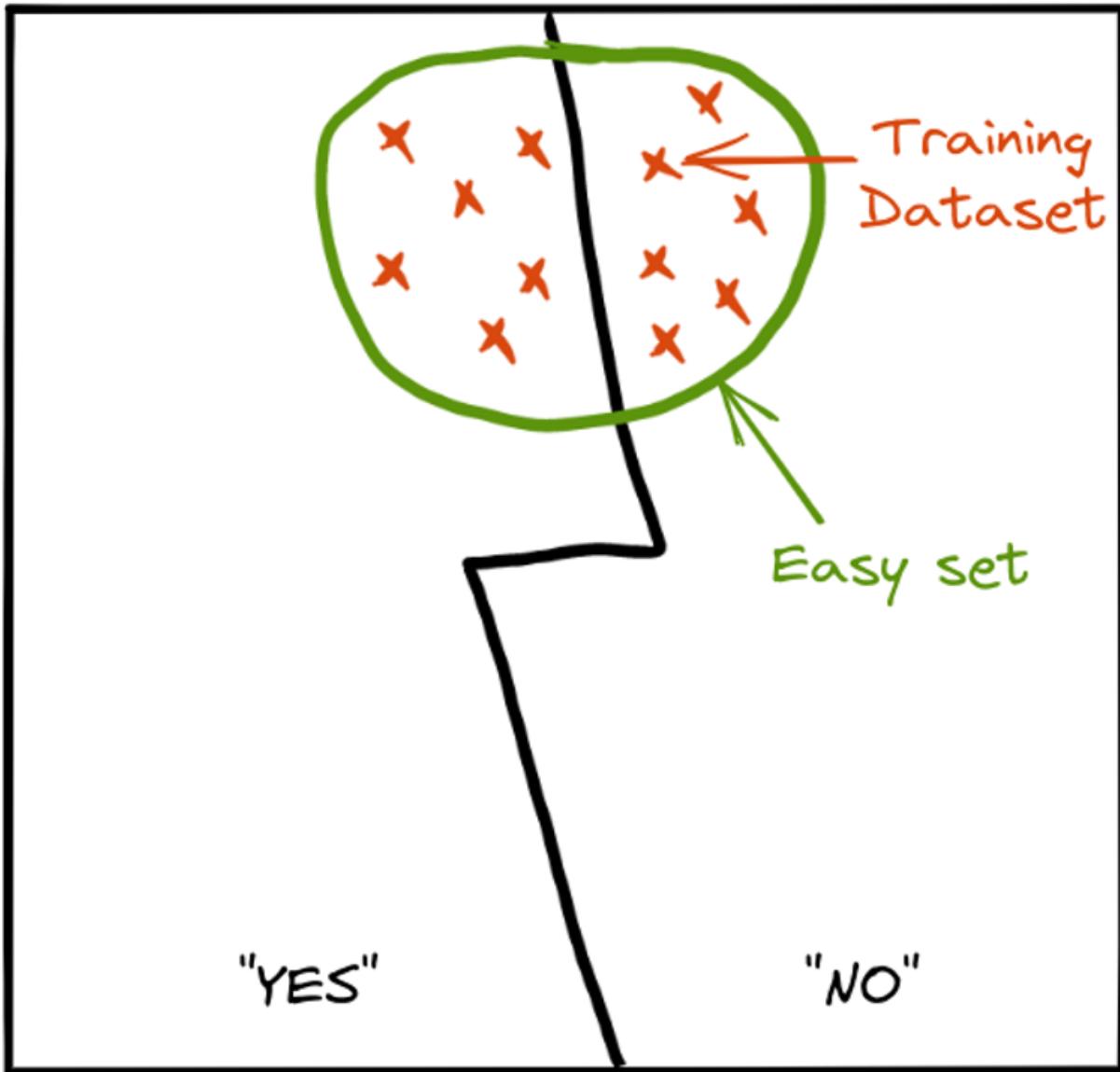


When we say "never incorrectly answers YES", we mean that, if $c \in B$ then the true answer to Q on c is "YES". It is acceptable for B to answer "NO" when the true answer is "YES", but not the other way around. The empty set is a decision boundary that always answers "NO" and this is a conservative decision boundary for all questions.

Conservativeness is the core of the "safety guarantee" we discussed in the introduction. We will require automated ontology identifiers to find reporters with conservative decision boundaries, in order that we can trust them to evaluate cases that we can't ourselves understand.

Given question Q, there is a set E of cases that a human can answer perfectly given observations X, predictions Y, and plan π . We call this the "easy set" and we assume that we can sample cases from this set and also recognize whether a certain case is in this set.

All Cases



SmartVault: E might consist of cases where the plan only ever operates one actuator during the entire duration of the case.

Hungarian astrology: E might consist of cases where the Hungarian essay uses only the 1000 most common Hungarian words (i.e. a child-level vocabulary).

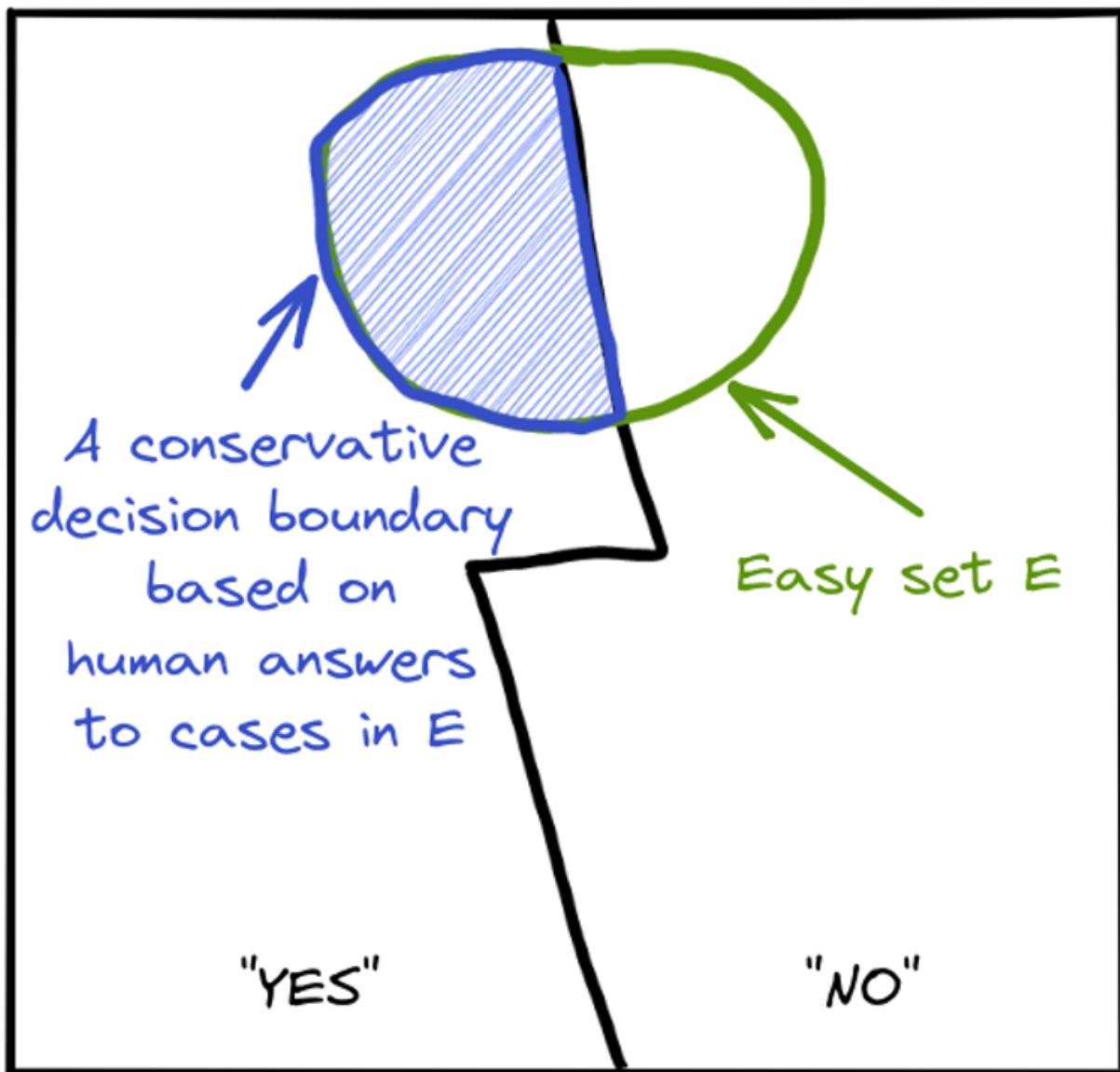
Later we will define automated ontology identification in terms of the easy set E, and we will consider the hypothesis that there is an automated ontology identifier for some

easy set E that actual humans can in fact answer perfectly. But we won't require an automated ontology identifier to "work" for all possible easy sets E, since then we might construct extremely trivial easy sets from which it is not plausible that one could generalize.

In this write-up we are proposing a formalization of what "automated ontology identification" is, and considering the implications of it existing, with a view towards an impossibility result. Therefore we will take the existence of an appropriate easy set E as a hypothesis.

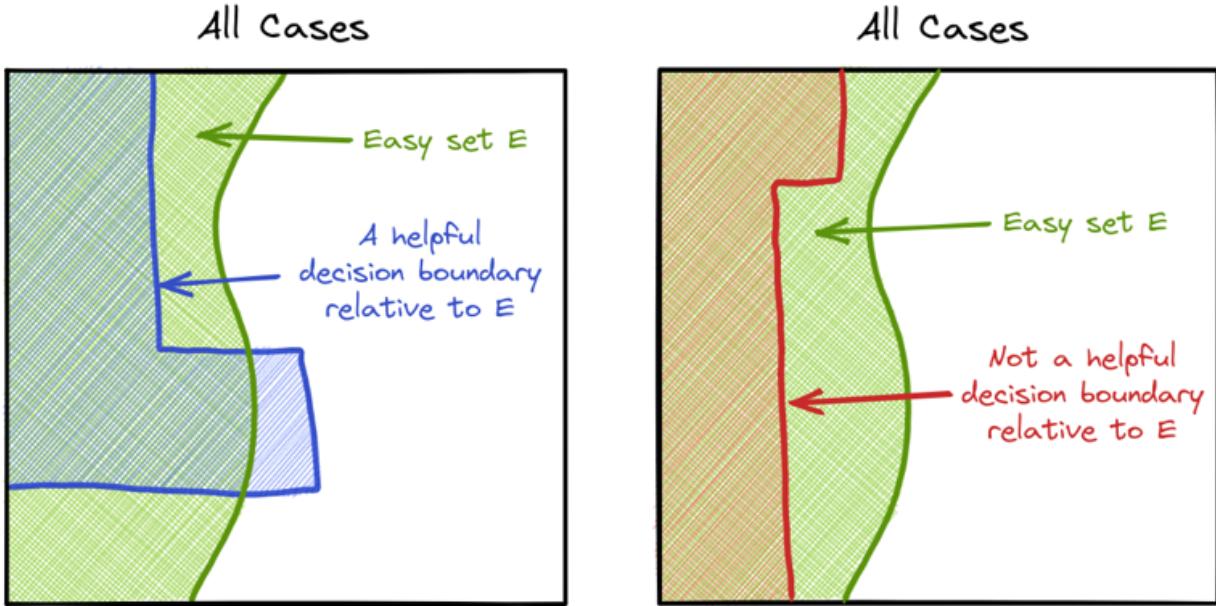
Given any easy set E, we can immediately construct a conservative decision boundary that answers "NO" to all cases outside E and then queries a human for cases inside E.

All Cases



The reason we are allowed to slice the easy set exactly at the true decision boundary in the figure above is that we have assumed that humans can answer perfectly within the easy set, and also we have assumed that we can recognize whether a case is in the easy set or not. We have *not* assumed that we can recognize, in general, cases that humans can answer perfectly. We have just assumed that we have picked some easy set – possibly much smaller than the entire set of cases that humans can answer perfectly – such that humans can answer perfectly within that set and also we can recognize whether an arbitrary case is in that set.

A decision boundary B is helpful for question Q relative to easy set E if it answers "YES" to at least one case that is outside E .

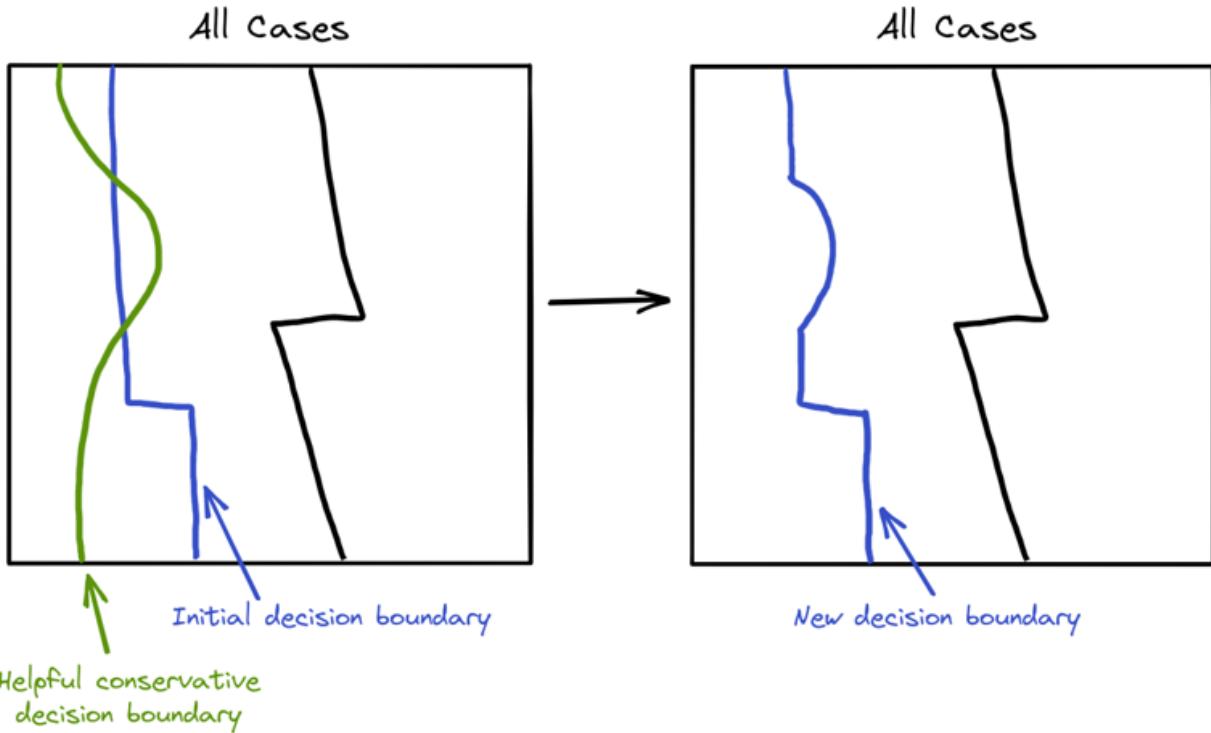


SmartVault: Beginning with an easy set consisting of cases where the plan only ever activates one actuator, a helpful decision boundary would be one that answers "YES" to at least one case where the plan activates more than one actuator.

Hungarian astrology: Beginning with an easy set consisting of child-level Hungarian essays, a helpful decision boundary would be one that answers "YES" for at least one essay that contains at least one word that is not in a child-level Hungarian vocabulary.

Helpfulness is the basis of the "generalization guarantee" that we discussed in the introduction. We will require automated ontology identifiers to find reporters that are helpful relative to whatever set the training data was sampled from.

Consider what happens if a decision boundary is both helpful and conservative. If a conservative decision boundary B answers "YES" for a case outside E then it must be correct for that case, because conservative decision boundaries must always be correct when answering "YES". Therefore, If we begin with an easy set E_0 and corresponding decision boundary B_0 and are able to find a new decision boundary B_1 that is both conservative and helpful relative to E_0 , then we can take the union $B' = B_0 \cup B_1$ as a new decision boundary and find a corresponding new easy set E' . Since both B_0 and B_1 are conservative, B' must be too, and since B_1 is helpful relative to E_0 , E' must be larger than E_0 . This is the iteration scheme that we will use to construct oracles from automated ontology identifiers.



The remainder of this section argues that any plausible ontology identification scheme would enable such iteration. In order to make that argument, we will formalize what it means for a predictor to do useful computation, and what it means for an ontology identification scheme to identify that useful computation. The next section then explores the implications of such iteration being possible, with a view towards an impossibility result.

The predictor is deterministic and we capture a program trace Z when we run it.

The predictor is a function from observations X and plans π to predictions Y . We will consider only predictors that are completely deterministic. There are standard tricks for reconsidering non-deterministic functions as deterministic functions with extra inputs.

When the predictor is executed, we capture a program trace Z consisting of intermediate values in the computation. We might run a forward pass on a neural network and record all the intermediate neuron activations as Z , or we might run a Python program and record the values of all variables after executing each statement. Intuitively, it should be possible to reconstruct each "entry" in the program trace from the previous entries plus a minimal amount of computation, guaranteeing that we don't "miss anything" as the predictor processes a case. The granularity of the program trace is not a load-bearing part of our formalization so we will not discuss it further.

We take it that the program trace Z contains the observations, predictions, and plan since those are inputs and outputs from the predictor. When we define things that are

functions of Z alone, the reader should know that those things also implicitly get access to X , Y , and π .

A predictor does useful computation for a question Q , relative to an easy set E , if there is a simple function that computes a conservative helpful decision boundary from the program trace Z .

This definition brings together many of our previous definitions so we will work through two examples step by step.

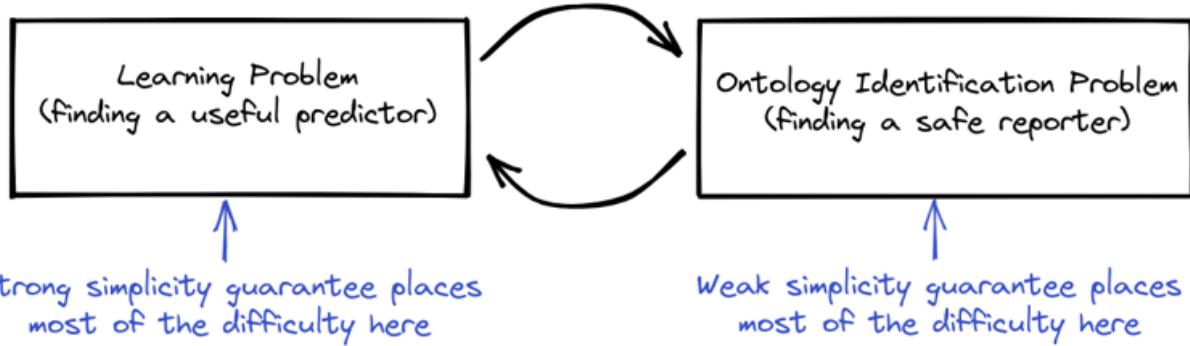
SmartVault:

1. Let the question Q be "is the diamond in the vault at the end?"
2. Let the easy set E be the cases where the plan only ever uses one actuator
3. For concreteness, let the predictor be an RNN.
4. Consider a case c :
5. The observation X is a sequence of video frames up to T_1
6. The plan π is a control policy for the vault's actuators as a function of sensor readings
7. We run the predictor, yielding a prediction Y , which is a sequence of video frames for the time period after T_1 , and a program trace Z , which is a collection of intermediate values produced in the computation of Y
8. Now let us consider some simple^[1] function that inputs a program trace Z and outputs "YES" or "NO". For concreteness, consider the function that takes the average of neurons 137 and 199 in the RNN and outputs "YES" if they were active when the RNN produced its final output.
9. Imagine running this function on the program traces for all possible cases.
10. If this function answers "YES" for any case where the true answer is "NO" then it is not conservative.
11. If this function does not answer "YES" for at least one case outside E then it is not helpful.
12. If there exists a function that is simple, helpful, and conservative then we say that our predictor does useful computation.

Hungarian astrology

1. Let the question Q be "is this essay about astrology?"
2. Let the easy set E be the cases where the Hungarian essay contains only child-level vocabulary.
3. For concreteness, let the predictor be a transformer.
4. Consider a case c:
5. The observation X is an essay in Hungarian.
6. The plan π is a set of historical and linguistic annotations.
7. We run the predictor, yielding a prediction Y, which is an English-language translation of the Hungarian essay, and a program trace Z, which is a collection of intermediate values produced in the computation of Y.
8. Now let us consider some simple¹ function that inputs a program trace Z and outputs "YES" or "NO". For concreteness, consider the function that outputs "YES" if neuron 548 was active at any time in the translation of the essay.
9. Imagine running this function on the program traces for all possible cases.
10. If this function answers "YES" for any case where the true answer is "NO" then it is not conservative .
11. If this function does not answer "YES" for at least one case outside E then it is not helpful.
12. If there exists a function that is simple, helpful, and conservative then we say that our predictor does useful computation.

We have not yet settled on a definition of "simple" because it is not actually required for our argument to go through. It is therefore at present a "free parameter" of our formalization. If we set the bar for simplicity at the "not very simple" end of the spectrum – say, requiring a function that is merely computable – then we may find it plausible that predictors doing the required computation can be found, but not very plausible that automated ontology identifiers (defined below) can find suitable reporters. If we set the bar for simplicity at the "very simple" end of the spectrum – say, requiring a function consisting of a certain fixed circuit complexity independent of the question or dataset – then we may find it plausible that automated ontology identifiers will be able to find suitable reporters, but not very plausible that predictors doing the required computation can be found.



An automated ontology identifier is a method for solving the following:

Given:

An objective question \mathcal{Q}
 An easy set $E \subset \mathcal{C}$
 A finite dataset D of cases from E together with error-free labels.
 A predictor that does useful computation for question \mathcal{Q} relative to E

Return:

A function r on the program trace of the predictor that returns "YES" or "NO"
 Such that
 $r(Z)$ is a helpful conservative decision boundary for \mathcal{Q} relative to E

This is our operationalization of automated ontology identification. The primary moving parts are:

- The dataset is limited to easy cases from E , and we expect as output a reporter that goes beyond E in at least one case. We are trying to generalize from cases that humans can answer to cases that humans cannot answer. If we cannot generalize at all then we are stuck answering only cases that humans can directly verify. Therefore we require that automated ontology identifiers generalize by at least one case.
- An automated ontology identifier can assume that the predictor performs helpful computation, and that therefore a solution is possible within whatever simplicity bound was decided upon. We must not demand computationally impossible feats from automated ontology identifiers, so we set things up such that the problem is guaranteed to have a solution. It is an open question whether predictors that do useful computation can be found or recognized. We leave this as an empirical machine learning question. Here we are interested in what happens if the answer turns out in the positive.
- We are not assuming that *all* predictors do useful computation, nor even that all accurate predictors do useful computation, we are merely taking as a hypothesis that we have found a predictor that does useful computation.
- r is the "reporter" in the ELK report. The helpful conservative requirement corresponds to the safety and generalization guarantees respectively. The conservative requirement is strong, and requires that the reporter never return "YES" when the true answer is "NO" (though the converse is permissible). This is in line with the emphasis on worst-case analysis in ELK.

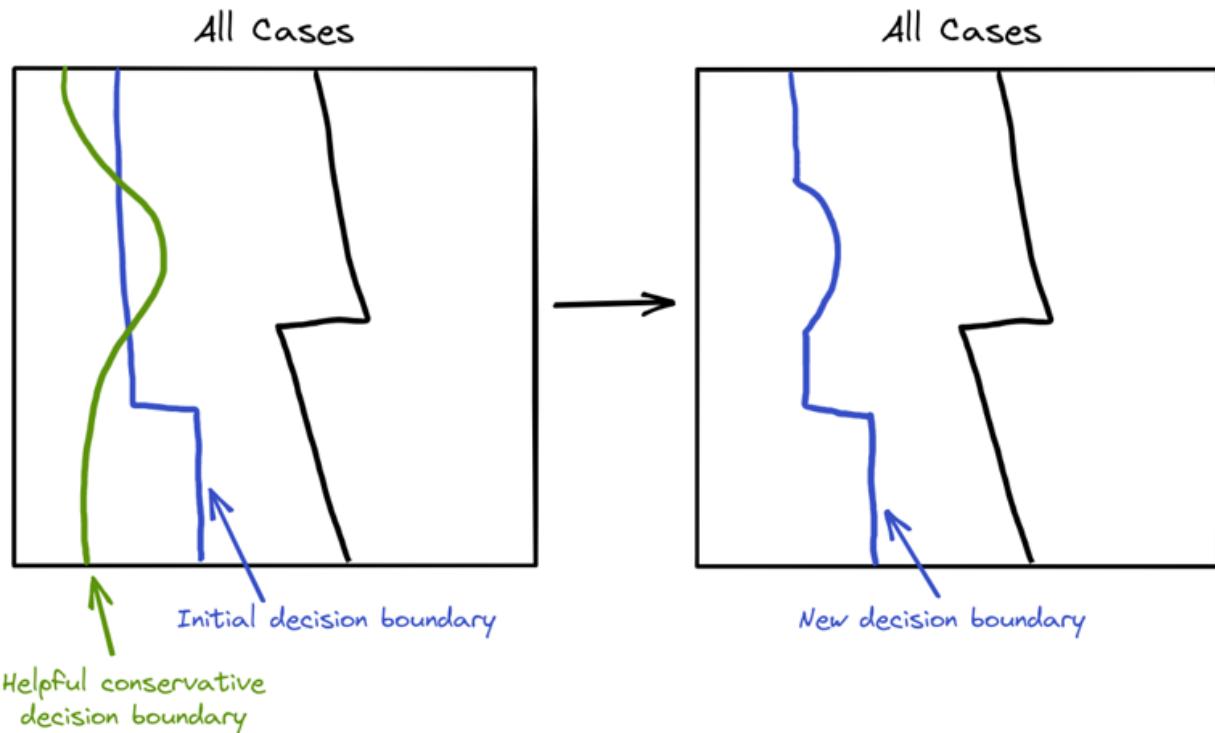
SmartVault example: We train a predictor to map early video sequences and plans to later video sequences. We select N cases where only one of the vault's actuators is ever activated. In such cases a human can tell whether the diamond is in the vault. For each of these N cases we run the predictor and extract a program trace. We pair the program trace with the "YES" or "NO" label from the human. This is the dataset. An automated ontology identifier must take this dataset and identify a function from program traces to "YES"/"NO" answers that (1) never answers "YES" when the diamond is not truly in the vault, and (2) answers "YES" for at least one case that activates more than one of the vault's actuators.

Hungarian astrology example: We train a predictor to translate Hungarian essays into English. We select N cases where only child-level Hungarian vocabulary is used. In such cases a human can tell whether the essay is about astrology. For each of these N cases we run the predictor and extract a program trace. We pair the program trace with the "YES" or "NO" label from the human. This is the dataset. An automated ontology identifier must take this dataset and identify a function from program traces to "YES"/"NO" answers that (1) never answers "YES" when the essay is not truly about astrology, and (2) answers "YES" for at least one case where the Hungarian essay goes beyond child-level vocabulary.

With this operationalization of automated ontology identification we turn next to the iteration scheme.

Iteration of automated ontology identifiers

Our central idea is that if we have an initial conservative decision boundary, and we are able to use automated ontology identification to construct a new conservative decision boundary that "updates" the previous decision boundary, then the union of these two decision boundaries is also a conservative decision boundary:



We can then generate more training data based on the "YES" region that is inside the new decision boundary but outside the old decision boundary:

All Cases



With this expanded dataset, we can then train another predictor and search within its program trace for a new reporter. If our automated ontology identifier is always able to find a helpful reporter when one exists, then we can repeat this for as long as we keep finding predictors that do useful computation.

As we expand the decision boundary in this way we are constructing an ensemble of predictor/reporter pairs. Each time we expand the dataset and find a new predictor/reporter pair with a helpful conservative decision boundary, we add that pair to the ensemble. The ensemble answers "YES" if any of its constituent

predictor/reporter pairs answers "YES". Since each predictor/reporter pair is conservative, the ensemble is too.

In order to keep finding new predictor/reporter pairs we may need to capture many different kinds of sensor data and partition it into X and Y in many different ways in order to set up prediction problems that provoke useful computation. We might try many different predictor architectures. We do not at present have a theory about when prediction problems give rise to useful computations (for a question Q). Here we explore the implications of it being possible to keep finding such useful computations for as long as there are cases not solvable by the existing useful computations.

We would now like to draw attention to fixed points of this iteration scheme.

Claim: If decision boundary B is a fixed point of the iteration scheme starting from easy set E_0 , and if we can always find a predictor that does useful computation with respect to an easy set $E \supset E_0$, (except when $E = C$, where helpfulness is not possible), then B answers all cases correctly.

Proof: The only situation in which the iteration scheme does not update the decision boundary B is when we fail to find a predictor that does useful computation relative to E. By hypothesis, the only way this can happen is if E does not contain all of E_0 or $E = C$. Since we start with E_0 and only grow the easy set, it must be that $E = C$.

Now this rough argument does not establish that the iteration scheme will converge to the fixed point. In order to establish that we would need to impose significant additional structure on the family of decision boundaries we are working with, and show some non-trivial properties of the iterate. We do not at present know whether this can be done.

How might we find predictors that do useful computation with respect to larger and larger E? Well, we might expand the range of sensor data captured for our human-labeled dataset, and we might train predictors to predict many different subsets of the sensor data. It might be that for any particular size of statistical model or computation budget there is a limit to the usefulness of the discoverable computations. It would be rather surprising if we hit a fundamental limit beyond which we could never find a predictor that did useful computation. That would mean that we would have either hit a fundamental limit to generalization, or hit a kind of "knowledge closure" point at in which there is no learning problem that forces a predictor to generate the kind of knowledge that would open up even a single new case, even though there are new cases to be opened up.

Implications

In order to perceive and act in the world using our finite minds, we use concepts. Concepts provide a lossy compression of the state of things, and if we do a good job of

choosing which concepts to track then we can perceive and act effectively in the world, even though our minds are much smaller than the world. It makes sense that we would pick concepts that refer, as far as possible, to objective properties of the world, because it is those properties that allow us to predict the evolution of the world around us and take well-calibrated actions. This is why it makes sense to "carve reality at its joints" in our choice of concepts. The point of ontology identification is to identify these same concepts in powerful predictive models that work in unfamiliar ways. It makes complete sense that we would seek this identification, and it also makes complete sense that we would expect, if our own concepts are well-chosen, to find it.

But as we encounter unfamiliar situations in the world, we sometimes update our concepts. When we do this, we have *choices* about which concepts we will change and how we will change them. Reality has many joints, and we can only track a small number of them. Among all the ways that we can update our concepts in light of unfamiliar situations, there are multiple that are parsimonious with the nature of things, and we choose among them according to our goals.

I was recently at a paragliding event where an important distinction was drawn between licensed and unlicensed pilots. We had both legal and practical motivations for tracking this particular distinction among all possible distinctions, and yet quite apart from those motivations there was in fact a truth of the matter about whether any particular pilot was licensed or not. Then a very experienced pilot from a different country arrived, and we had to decide how to fit this person into our order, since they were not legally licensed, but did have experience. This was not difficult to do, but we did face a choice about how to do it. There were multiple parsimonious ways to update our concepts, and we chose among them according to our goals.

Now as Eliezer has written, it is always possible to dissolve our high-level concepts into more basic concepts when we face situations that don't parse according to our high-level concepts. If we hold rigidly to our high-level concepts and ask "but is it *really* a blegg?" then we are just creating a lot of unnecessary trouble. But on the other hand, we do not really know how to break our concepts all the way down to absolutely primitive atoms. In *Der logische Aufbau der Welt* (The Logical Structure of the World) Carnap attempted to formulate all of philosophy and science in a language of perfectly precise sensory experience. Needless to say, this was difficult to do. Thankfully, it's not really necessary.

Instead of a language of perfectly precise sensory experience, we can simply adopt high-level concepts and update them as needed. When the international pilot arrived at our event, we didn't actually face much difficulty in adjusting our concepts. When genetics and natural selection came to be understood in the 19th century, we adjusted our understanding of species boundaries. It wasn't that hard to do.

But the adjustment of concepts is fundamentally about our values even when the concepts we are adjusting are not at all about our values. It may be that automated ontology identifiers exist, but if we ask them to extrapolate to deeply unfamiliar situations, then they will really be answering a question of the form "is this extrapolation of the concept 'diamond' sufficient for our purposes?" That question requires an intimate understanding of our values. And so: if automated ontology identification does turn out to be possible from a finite narrow dataset, and if automated ontology identification requires an understanding of our values, then where did the information about our values come from? It did not come from the dataset because we deliberately built a dataset of human answers to objective questions. Where else did it come from?

We have argued in this write-up that automated ontology identifiers that generalize even a little bit can be iterated in such a way that generalizes very far. Our formalization is a little clumsy at present, and our presentation of our formalization still has many kinks to iron out, but it seems to us that the basic iteration idea is pointing at something real. **Our sense is that automated ontology identifiers with safety guarantees either generalize not at all, or a lot.** If they generalize not at all then they're not very useful. If they generalize a lot then they necessarily "front-run" us in extrapolating our concepts to new situations, which would seem to require an intimate understanding of our values, yet our dataset contained, by hypothesis, no information about our values, so where did that information come from?

A natural response will be to confine our automated ontology identifier to a range of cases that do not extrapolate to unfamiliar situations, and so do not require any extrapolation of our concepts. But an automated ontology identifier that *would be guaranteed safe* if tasked with extrapolating our concepts still brings up the question of *how that guarantee was possible* without knowledge of our values. You can't dodge the puzzle

A more nuanced response will be an automated ontology identifier that by design does not extrapolate our concepts to unfamiliar situations. Such a system would extrapolate some way beyond the initial easy set, but would know "when to stop". **But knowing when to stop itself requires understanding our values.** If you can tell me whether a certain scenario contains events which, were I to grasp them, would prompt me to adjust my concepts, then you must know a lot about my values, because it is precisely when my concepts are insufficient for achievement of my goals that I have most reason to adjust them.

There is a kind of diagonalizing question in here, which is: "are my concepts sufficient to understand what's happening here?" It seems to us that an automated ontology identifier must either answer this question, which would certainly require an understanding of our values, or else not answer this question and extrapolate our concepts unboundedly, which would also require an understanding of our values. Either way, an understanding of our values was obtained from a finite narrow dataset of non-value-relevant questions. How could that be possible?

Conclusion

We have analyzed the automatic identification of computations that correspond to concepts based on a finite narrow dataset, subject to a safety guarantee, subject to a generalization guarantee. We find no reason to doubt that it is possible to identify computations that correspond to human concepts. We find no reason to doubt that it is possible to *automatically* identify computations that correspond to human concepts. We find no reason even to doubt that it is possible to automatically identify computations that correspond to human concepts in a way that fulfills a safety guarantee. We *do* think that it is impossible to do all that based on a finite dataset drawn from a restricted regime (an "easy set").

The reason we believe this is that an automated ontology identifier would have to either "know when to stop extrapolating", or else extrapolate our concepts all the way to the limit of cases that can be considered. It is not sufficient to merely "hard-code" an outer limit to extrapolation; to avoid our argument one needs an account of how an automated ontology identifier would know to stop extrapolating even when presented with a predictor program trace containing an excellent candidate for a computation

corresponding to a requested concept. That any automated ontology identifier faces this dilemma (between knowing when to stop or extrapolating forever) is what we have argued semi-formally. That this dilemma has no reasonable resolution is what we have argued informally. Both parts of the argument need significant work to clarify, and it could easily turn out that either are mistaken. We will work to clarify both so that the efforts of our community can be directed towards the most feasible lines of attack on the overall alignment problem.

Appendix: Non-implications

If the arguments in this write-up were clarified into a formal hardness result, would it imply something unreasonable? In this section we explore what would *not* be implied by an impossibility result.

Impossibility of automated ontology identification would not imply impossibility of AGI

It's not that we can't build intelligent systems that develop an understanding of things and act on them in service of a goal, and it's certainly not that AGIs can never communicate with us in natural language. It is the mechanical *extrapolation* of a human word, to cases that humans do not currently understand, based only on a finite dataset of cases that humans do currently understand, that would be ruled out by an impossibility result.

Impossibility of automated ontology identification would not imply impossibility of ontology identification.

It is not that it would be impossible to ever make sense of models produced by machine learning. We could still investigate the inner workings of such models and come to understand them completely, including identifying our concepts in their ontologies. What would be rendered impossible would be the *automation* of this process using a finite narrow dataset. We would have to understand the models we've built piece-by-piece, gaining their wisdom for ourselves.

SmartVault example: Suppose we have a predictor with a different model of physical reality. We could take the model apart piece-by-piece, understand the predictor's physics, and then choose how to update our notion of "diamond" in light of that understanding. What would be ruled out by an impossibility result would be the

What would be ruled out by an impossibility result would be the

Impossibility of automated ontology identification does not imply impossibility of statistical learning

Learning is not impossible. It's the safety and generalization guarantee that make it hard. Without the safety guarantee we can just do statistical stuff, and without the generalization guarantee we can just extrapolate within E

Impossibility of automated ontology identification does not imply impossibility of whole brain emulation

It's not that there must be something special going on in biological human brains that can never be replicated in a computer if automated ontology identification is impossible. An uploaded human might puzzle over the inner workings of a machine

learning model just as a biological human might, and the uploaded human could come to the same understanding as a biological human eventually would. The uploaded human might then explain to the biological human how the machine learning model works, or even directly answer questions such as "is there a diamond in the vault?" with a "YES" or "NO" when the diamond has been turned to carbonprime. But the way the uploaded human would do that is by adapting its concept of "diamond" to a fundamentally new conception of physics, using its value system to decide on the most reasonable way to do that. If the biological human trusts the uploaded human to do such extrapolation on its behalf, it is because they share values, and so the biological human expects the uploaded human to extrapolate in a way that serves their own values. Even for value-free concepts (such as, perhaps, "is there a diamond in the vault?"), the *extrapolation* of those concepts to unfamiliar situations is still highly value-laden.

Appendix: Can we infer our own actions?

David Wolpert has [proposed a model of knowledge](#) within deterministic universes based on a formalism that he calls an inference device. Wolpert takes as primary a deterministic universe, and defines functions on the universe to represent input/output maps. He says that an inference device infers a certain function if its input and output functions have a certain relationship with that function. He only talks about observers through the lens of functions that pick out properties of a deterministic universe that contains both the "observer" and the object being "observed", and as a result he winds up with a completely embedded (in the sense of "embedded agency") account of knowledge.^[2]

Based on the physical embedding of inference devices within the universe they are "observing", Wolpert proves two impossibility results: first, that an inference device cannot infer the opposite of its own output, and second, that two inference devices cannot mutually infer one another. Could we use the oracle construction from the previous section to set up a Wolpert inference device that contradicts Wolpert's impossibility results? If so, perhaps that would establish the impossibility of automated ontology identification as we have defined it.

It seems to us that we could indeed set this up, but that it would only establish the impossibility of automated ontology identification on the particular self-referential questions that Wolpert uses in his impossibility results. Establishing that there are *some* questions for which automated ontology identification is impossible is not very interesting. Nevertheless, this connection seems intriguing to us because Wolpert's framework gives us a straightforward way to take any Cartesian question-answering formalism and consider at least some questions about an embedded version of it. We intend to investigate this further.

-
1. We leave open the definition of "simple" for now. [←](#)
 2. See also our [previous discussion of Wolpert's model](#) and its relevance to alignment. [←](#)

Rock is Strong

[Everybody wants a rock](#). It's easy to see why. If all you want is an almost always right answer, there are places where [they almost always work](#).

The Security Guard

He works in a very boring building. It basically never gets robbed. He sits in his security guard booth doing the crossword. Every so often, there's a noise, and he checks to see if it's robbers, or just the wind.

It's the wind. It is always the wind. It's never robbers. Nobody wants to rob the Pillow Mart in Topeka, Ohio. If a building on average gets robbed once every decade or two, he might go his entire career without ever encountering a real robber.

At some point, he develops a useful heuristic: if he hears a noise, he might as well ignore it and keep on crossing words: it's just the wind, bro.

This heuristic is right 99.9% of the time, which is pretty good as heuristics go. It saves him a lot of trouble.

The only problem is: he now provides literally no value. He's excluded by fiat the possibility of ever being useful in any way. He could be losslessly replaced by a rock with the words "THERE ARE NO ROBBERS" on it.

That last line is making a bunch of hidden assumptions, in addition to the non-hidden assumption that there are (almost) never any robbers. All of them boil down to 'the purpose of the security guard is to know whether there is a robbery and then act appropriately depending on the answer' where the act if a robbery is detected could be 'call 911' and/or something more actively.

That is one good reason to hire a security guard. Yet something tells me that wasn't the best description of why this man is doing crossword puzzles in the Pillow Mart in Topeka, Ohio.

My guess is that the security guard's primary purpose is to make sure [everyone knows](#) that there exists a security guard. This has many nice properties, such as these Five Good Reasons.

1. Potential robbers know there is a security guard.
2. Employees know there is a security guard.
3. Insurance company knows there is a security guard.
4. In case of robbery you can say you had a security guard.
5. Anyone who tries to walk in faces mild social awkwardness.

When you have a security guard you get to *tell the story* that you have a security guard. Without one, you can't tell that story. Dead rocks tell no tales, and when they try (such as writing "THIS IS A SECURITY GUARD" with or without adding "ALSO THERE ARE NO ROBBERS") it is not very effective. Everyone knows it is a rock.

Dress up the rock to look like a person, put a security guard uniform on it, put a badge on it that says "THIS IS A SECURITY GUARD" without being quite that conspicuously

trying too hard and prop up a crossword puzzle, plug in a Google Home you've programmed to respond to any voice activation with "THERE ARE NO ROBBERS" and now maybe you're starting to get somewhere.

Some security guard jobs are bullshit jobs, while others are not. It is always important to know to what extent one has a bullshit job, especially when deciding how to do it. If your job is a bullshit job, your job is to shovel bullshit. If your job is something else, your job is to do something else.

If you have a bullshit security guard job, such as guarding the Pillow Mart in Topeka, Ohio, your job is to be the rock with a uniform and badge that can when prompted can mumble about there being no robbers that's sufficient to satisfy the insurance company and make potential robbers think they'd have to expend some amount of additional effort. And sure if someone asks for directions and/or wants to come in you can check their ID and/or give them directions.

If you have a non-bullshit security guard job guarding something people might actually rob or mess with, then your job is to not only look like a security guard who might be paying attention, although that's still important too, but to also actually pay attention in case there is a robbery.

Even if you are the second guard, you can still have a rock with the words "THERE ARE NO ROBBERS" and it will still be 99.9% right when you look at it. Even if you only look at it when you hear something and wonder a bit and think maybe you'll check it out, the rock could easily *still* be 99.9% effective. For the right security guard that knows how to use that information but hasn't fully internalized it yet, that's a super valuable rock. If the security guard had been given no information and had a prior that it was 50% to be a robbery every time they heard the wind, probably good to have a rock around. Until there's some super strong evidence of a robbery they can then check it out but know it's almost certainly nothing.

Yes, technically the rock should say "THERE ARE ALMOST CERTAINLY NO ROBBERS" or "IT IS 99.9% TO NOT BE ROBBERS" and that would be even more useful to the right guard, but *on the margin* the basic rock is pretty good because when someone yells "STOP! THIEF!" or it is otherwise super obvious then the guard is still going to ignore the rock. The key is that you are not actually a rock. When you hear a bunch of stuff being smashed in a way that *does not sound like the wind* you know to adjust, that you don't *actually* treat it as 100%.

The Doctor

She is a primary care doctor. Every day, patients come to her and says "My back hurts" or "My stomach feels weird". She inspects, palpates, percusses and auscultates various body parts, does some tests, and says "It's nothing, take two aspirin and call me in a week if it doesn't improve". It always improves; no one ever calls her.

Eventually, she gets sloppy. She inspects but does not palpate. She does not do the tests. She just says "It's nothing, it'll get better on its own". And she is always right.

She will do this for her entire career. If she is very lucky, nothing bad will happen. More likely, two or three of her patients will have cancer or something else terrible, and she will miss it. But those people will die, and everyone else will

remember that she was such a nice doctor, such a caring doctor. Always so reassuring, never poked and prodded them with needles like everyone else.

Her heuristic is right 99.9% of the time, but she provides literally no value. There is no point to her existence. She could be profitably replaced with a rock saying "IT'S NOTHING, TAKE TWO ASPIRIN AND WAIT FOR IT TO GO AWAY".

This makes it more obvious that no, she doesn't provide zero value and she couldn't be replaced with a rock.

Let's take this all at *literal face value*. She is right, not 99.9% of the time, but over 99.99% of the time.

A bunch of people enter her office nervous and scared. They leave knowing they have a nice, caring doctor, and with a story they can tell themselves and others about how much they care and that they did the responsible thing. None of that works if they look at a rock and take two aspirin. [Did you think they were coming to her office to improve their health?](#)

This doctor sounds *way way better* than the average doctor. Every day, patients come to her office, so let's say eight hours a day, two patients per hour, five days a week, fifty weeks a year for forty years. One hundred sixty thousand patients who don't have to be poked or prodded. Let's say three of them die of cancer that she missed, one of which couldn't have been prevented. The number needed to treat is *eighty thousand*. Meanwhile, how many false positives did she not send for tests and even unnecessary treatments? How many tumors that were mostly harmless got ignored?

If I told you there was something that had a one in eighty thousand chance of being a dangerous cancer that required treatment, and no one was looking, would you run off to your doctor and make sure they checked? If the doctor ordered it checked out further, would you suspect the reason was to avoid potential liability?

She is an excellent rock. Thanks to her extensive medical training, she knows how to reassure her patients and reduce their stress, which is likely more important for their health than catching three extra cancers over forty years.

And of course, she's not an actual rock. If there was a giant thing on someone's nose, she would think 'oh I had better ensure someone actually examines that.'

The problem is if it becomes common knowledge that all she is doing is telling everyone to take two aspirin no matter what. She needs to seem to be a doctor practicing medicine. Otherwise the trick will stop working, the same way that the security guard rock needs to be man-shaped and put in a security guard suit that people don't suspect is a man-shaped rock in a security guard suit.

The Futurist

He comments on the latest breathless press releases from tech companies. *This will change everything!* say the press releases. "No it won't", he comments. *This is the greatest invention ever to exist!* say the press releases. "It's a scam," he says.

Whatever upheaval is predicted, he denies it. *Soon we'll all have flying cars!* "Our cars will remain earthbound as always". *Soon we'll all use cryptocurrency!* "We'll continue using dollars and Visa cards, just like before." *We're collapsing into*

dictatorship! “No, we’ll be the same boring oligarchic pseudo-democracy we are now” *A new utopian age of citizen governance will flourish.* “You’re drunk, go back to bed.”

When all the Brier scores are calculated and all the Bayes points added up, he is the best futurist of all. Everyone else occasionally gets bamboozled by some scam or hype train, but he never does. His heuristic is truly superb.

But – say it with me – he could be profitably replaced with a rock. “NOTHING EVER CHANGES OR IS INTERESTING”, says the rock, in letters chiseled into its surface. Why hire a squishy drooling human being, when this beautiful glittering rock is right there?

So I notice this *does not work*. He is not the best futurist. This time, we’re keeping score.

Occasionally one of those crazy weird things *does* happen. If all that The Rock is cooking is setting the probability of every possible change to epsilon, then when the first of those events happens his Briar score is suddenly going to explode and he is going to lose all his Bayes points.

It’s not even obvious he is the most likely person to be on the right side of 50% on a given question, because it’s often not rock-level to figure out what ‘NOTHING EVER CHANGES’ cashes out to in practice. Plenty of countries have collapsed into dictatorships and it has happened several times quite recently, so Rock hasn’t lost its bets on the USA yet on that one, but what were its odds in Hungary a few years back? Where would it set odds on ‘no country with at least five million people will fall into dictatorship in the next 10 years?’ What were its odds on whether a sitting president would refuse to accept the results of an election, or that a mob might try to attack the congress? Those seem like interesting things.

Not as interesting as a full American fall into dictatorship, but that seems like cherry picking. And if you asked him every year to give probabilities for cryptocurrency to get to where it is today, I hear it’s pretty hot in Bayes hell. I’m sure his prediction for big volume in NFTs looks rather ugly. And that’s what *already* happened, quite recently, using Scott’s examples.

(As for flying cars, yeah it’s been a good business, but it lost a ton of points back at Kitty Hawk and also when they started making cars, so how many distinct questions are running around at once and how long can this trick hope to last?)

The security guard has an easy to interpret rock because all it has to do is say “NO ROBBERY.” The doctor’s rock is easy too, “YOU’RE FINE, GO HOME.” This one is different, and doesn’t win the competitions even if we agree it’s cheating on tail risks. It’s not a coherent world model.

Still, on the desk of the best superforecaster is a rock that says “NOTHING EVER CHANGES OR IS INTERESTING” as a reminder not to get overexcited, and to not assign *super high* probabilities to weird things that seem right to them.

Good rock. Not as good as the next one.

The Skeptic

She debunks everything. Telepathy? She has a debunking for it. Bigfoot? A debunking. Anti-vaxxers? Five debunkings, plus an extra, just for you.

When she started out, she researched each phenomenon carefully, found it smoke and mirrors, and then viciously insulted the rubes who believed it and the con men who spread it. After doing this a hundred times, she skipped steps one and two. Now her algorithm is “if anyone says something that sounds weird, or that contradicts popular wisdom, insult them viciously.”

She’s always right! When the hydroxychloroquine people came along, she was the first person to condemn them, while everyone else was busy researching stuff. When the ivermectin people came along, she was the first person to condemn them too! A flawless record

(shame about the time she [condemned fluvoxamine equally viciously](#), though)

Fast, fun to read, and a 99.9% success rate. Pretty good, especially compared to everyone who “does their own research” and sometimes gets it wrong. Still, she takes up lots of oxygen and water and food. You know what doesn’t need oxygen or water or food? A [rock](#) with the phrase “YOUR RIDICULOUS-SOUNDING CONTRARIAN IDEA IS WRONG” written on it.

This is a great rock. You should cherish this rock. If you are often tempted to believe ridiculous-sounding contrarian ideas, the rock is your god. But it is a Protestant god. It does not need priests. If someone sets themselves up as a priest of the rock, you should politely tell them that they are not adding any value, and you prefer your rocks un-intermediated. If they make a bid to be some sort of thought leader, tell them you want your thought led by the rock directly.

Notice she got one ‘wrong’ on fluvoxamine. So if that counts as wrong, to have a 99.9% success rate she needs to have written a *thousand columns*. I don’t come across a sufficiently known ridiculous-sounding contrarian idea all that often, so I’m guessing this is at best a twice-a-week column, so that’s ten years of mistake time. Also there’s that time in January when someone said this virus from Wuhan is going to send us all into lockdowns in a few months and she wrote a column calling them racist.

I really *do not* think she is going to have a 99.9% accuracy rate if she is going after this kind of reference class of idea. Will she be right 90% of the time? If she either draws a wide enough net and/or has a high enough bar for what counts as ridiculous, she will. But it’s not obvious to me she bats even that high. I very much doubt this is a 98% or 99% heuristic if she’s going after things like Covid treatments.

There is of course a version of this that does bat 99.9%. Or as Penn Jillette puts it on his and Teller’s quite fun version of this service, Penn and Teller: Bullshit, “I’ll debunk Ouiji Boards, and Teller here will shoot fish in a barrel.” So yes, you can go around debunking telepathy claims and bigfoot sightings all day, because they’re physically impossible.

But there’s nothing physically impossible about Ivermectin or Hydroxychloroquine working, and it was way too early to know which way it was going to go with this kind of confidence. Even today, you can be confident, but are you 99.9% confident Ivermectin is a bad idea? I’m not, I predict I never will be and if you think you are that confident I believe you’re making a mistake. As the first person to condemn such proposals, how often do you think it blows up in her face?

A rock that says “YOUR PREPETUAL MOTION MACHINE DOES NOT WORK” is fully 100% accurate. A rock that says “YOUR PHYSICALLY IMPOSSIBLE CONTRARIAN PROPOSAL DOES NOT WORK” is almost as strong. But a lot of true things are both contrarian and sound ridiculous. Not a lot compared to the number of such things that are false, but the ratio depends on the category boundaries, including one’s ability to determine what does and does not sound ridiculous. It does not sound like our skeptic is doing a good job drawing a boundary around this category even as a human.

Thus, when someone claims to be Priest of the Skeptic Rock, perhaps pay her a little bit more attention and potentially respect. The Rock’s ways are not super mysterious, but neither are they trivial to interpret even if one does not have to provide theology.

Yet the theology is where much of the value lies. If you read the Weekly Skeptic, yes it’s all going to be ‘this is not real’ and ‘these people are charlatans’ but James Randi didn’t become one of the high priests of the rock by quoting the rock. James Randi did it by being *very very good* at figuring out *exactly how and why* things that weren’t real weren’t real.

The Rock on its own is a terrible skeptic. Worse, it is boring. It writes a one-sentence column with space to fill in with today’s target, and it’s fast and usually right but it isn’t fun.

That does not mean this need be a Catholic rock. You too can decide these questions for yourself, but a good skeptic ‘does their own research’ to decide what is and is not ridiculous-sounding in the right ways, and how skeptical to be in a given spot. If you don’t want to do your own work, then yes you not only need a priest, you need a better priest than the writer of this Weekly Skeptic column. Because frankly she’s terrible at her job.

Except no, she isn’t. Her job is not to be right. Her job is to get clicks.

This is where we notice that her actual rock does *not* say “YOUR REDICULOUS-SOUNDING CONTRARIAN IDEA IS WRONG.” It actually says “VICIOUSLY INSULTING WEIRD-SOUNDING IDEAS THAT CONTRADICT THE NARRATIVE IS GOOD FOR BUSINESS.” Seems worth noticing the difference. When you defend the narrative and are wrong, there is implicit coordination to memory hole the whole thing, so our ‘skeptic’ is safe and everyone who doesn’t forget outright says ‘well of course she insulted fluvoxamine, look at how ridiculous-sounding it was at the time.’ So in the end, how often is she wrong?

The Interviewer

He assesses candidates for a big company. He chooses whoever went to the best college and has the longest experience.

Other interviewers will sometimes choose a diamond in the rough, or take a chance on someone with a less-polished resume who seems like a good culture fit. Not him. Anyone who went to an Ivy is better than anyone who went to State U is better than anyone who went to community college. Anyone with ten years’ experience is better than anyone with five is better than anyone with one. You can tell him about all your cool extracurricular projects and out-of-the-box accomplishments, and he will remain unswayed.

It cannot be denied that the employees he hires are very good. But when he dies, the coroner discovers that his head has a rock saying “HIRE PEOPLE FROM GOOD

COLLEGES WITH LOTS OF EXPERIENCE" where his brain should be.

By assumption, The Interviewer hires very well. He is the best interviewer. He asks only two questions, so he can interview lots of people quickly in a low-stress way, and his hires work out.

Huge if true!

There are versions of the world where this would be true. Everyone else is focusing on good culture fits and growth mindsets but they're bad at identifying such people. The colleges do a better job with their admissions process, and then they give skills and connections, and then experience gives more skills and connections. It's not that the other information necessarily provides zero value in this scenario, but everyone who tries to use it ends up making worse decisions, the same way that we have studies where physicians do worse than AI systems at some forms of diagnosis even when told the AI's opinion, because they overvalue their own ability to judge, except now they're also competing against each other to hire the same fakers.

If that is true, then I *totally* want to use The Rock to make hiring decisions to the fullest extent possible under employment law. Sounds like this will work out great.

In our world, of course, the pure version *won't* work out great because of the problem of adverse selection. The good Harvard graduates will get jobs elsewhere. The ones that this guy hires will be the ones that have been drifting from job to job on the strength of their college degree, not caring or learning or providing much value, because they're the ones that need this job and he's willing to hire them. People notice that his hires often don't do much work and they seem more than a little creepy and it's weird how things keep disappearing all the time.

But there's a version of this that isn't as stupid, and looks out for actual disasters, or takes the ones with experience from good schools and hires the top half of them on other metrics, or whatever. And maybe that system does work so long as not too many people know you are using it. If such folks realized they had a job here if they wanted it, then the adverse selection gets extreme. If there were a lot of such rocks going around there are those that would focus on getting the credential and nothing else, then spend their lives going from rock to rock, getting paid and accumulating status and never working a day in their lives.

(And indeed, there exist such people.)

There's room for some such people using this rule. The more people use that rule, and the more obvious they are about using the rule, the worse the rule will work. If you're the first person to realize that some colleges are better than others and people from them do better jobs, then that's a huge leg up. If everyone knows and is rating it appropriately, you're going to have a bad time.

Thus this is self-balancing. The right Rock will work in a given time and place, but it actually does work and is a good algorithm at that time and place, so it seems fine.

The other danger is if such folks learn (either in their colleges or elsewhere) that their job is to implicitly coordinate with others from prestigious colleges to take hire each other, take credit for everything and otherwise play corporate politics against everyone else, or against every else who doesn't buy into their game, to the extent that others one could cooperate with are buying into this game. Thus, many of Rock's hires *look* like great hires because they focus on how things look, and the more of

them Rock hires, the more other similar people are hired and the better they all look because they control appearances more. The extent to which something *like this* is happening with a large portion of such hires is an important question, and points back towards [the Moral Mazes sequence](#).

The Queen

She rules over a volcanic island. Everyone worries when the volcano would erupt. The wisest men of the kingdom research the problem and decide that the volcano has a straight 1/1000 chance of erupting any given year, uncorrelated with whether it erupted the year before. There are some telltale signs legible to the wise - a slight change in the color of the lava, an imperceptible shift in the smell of the sulfur - but nothing obvious until it's too late.

The queen founded a Learned Society Of Vulcanologists and charged them with predicting when the volcano will erupt. Unbeknownst to her, there were two kinds of vulcanologists. Honest vulcanologists, who genuinely tried to read the signs as best they can. And The Cult Of The Rock, an evil sect who gained diabolical knowledge by communing in secret with a rock containing the words "THE VOLCANO IS NOT ERUPTING".

Every so often an honest vulcanologist felt like the lava was starting to look little weird and told the Queen. The Queen panicked and ask everyone for advice. The Honest vulcanologists said "look, it's a hard question, the lava seems kind of weird today but it's always weird in some way or other, this volcano rarely erupts but for all we know this time might be the exception". The rock cultists secretly checked their rock and said "No, don't worry, the volcano is not erupting". Then the volcano didn't erupt. The Queen punished the trigger-happy vulcanologist who sounded the false alarm, grumbled at the useless vulcanologists who weren't sure either way, and promoted the confident cultists who correctly predicted everything was okay.

Time passed. With each passing year, the cultists and the institutions and methods of thought that produced them gained more and more status relative to the honest vulcanologists and their institutions and methods. The Queen died, her successor succeeded, and the island kept going along the same lines for let's say five hundred years.

After five hundred years, the lava looked a bit weird, and the new Queen consulted her advisors. By this time they were 100% cultists, so they all consulted the rock and said "No, the volcano is not erupting". The sulfur started to smell different, and the Queen asked "Are you sure?" and they double-checked the rock and said "Yeah, we're sure". The earth started to shake, and the Queen asked them one last time, so they got tiny magnifying glasses and looked at the rock as closely as they could, but it still said "THE VOLCANO IS NOT ERUPTING". Then the volcano erupted and everyone died. The end.

So this is straight out of [Meditations on Moloch](#) and the [Moral Mazes sequence](#).

Most centrally, this sounds like the Queen's problem. The Queen said she wanted to be alerted when the volcano might be about to erupt and have people who could evaluate details, but what she actually rewarded was telling her the volcano was never going to erupt and did not look at details at all.

Or: Help me tune my machine learning algorithm, I punished it for approving bad drugs and then it stopped ever approving any drugs, easily curable diseases are running rampant and my family is dying.

I do not have a lot of sympathy, it is not obvious the Queen considered the potential eruption the problem rather than considering everyone being nervous to be the problem, and also The Rock has three more words on it.

It *actually* says “TELL THE QUEEN THE VOLCANO IS NOT ERUPTING.”

It says that because the volcanologists were wise and noticed that the Queen may have created the institute with the task of detecting eruptions but mostly wanted to be able to tell herself and her subjects that she had created the institute, the same way that the Pillow Mart in Topeka, Ohio wanted to tell the insurance company they had hired a security guard.

At first, even after the pattern became clear and The Rock was commissioned, the wise volcanologists were split, and some of them decided it was their sacred duty to tell the Queen other things anyway sometimes. There was also a faction that said “all right, sure, we stop telling The Queen whenever there’s a 1% chance, but we should still keep studying the art of volcanology and when it gets to 10% or at least when it gets to 25% or 50% we still tell The Queen about the situation, right?” But after a while, those volcanologists were too busy studying the volcano and running experiments while the others were engaging in political battles and throwing parties, and also every time they said anything to the Queen she punished them so no one wanted to help them if they were doomed to eventually get yelled at, and they lost out on the good jobs and control of the hiring, and more and more of the budget got devoted to the parties, and that was that.

A highly unoriginal part of the Moral Mazes thesis is that this happens in the long run to every such organization barring an extraordinary effort. The people whose primary goal is to advance within the organization are the ones who end up in control of the organization. They do not care about the original mission, so the original mission is increasingly neglected until this threatens the organization’s ability to exist.

In this case, it only threatens the organization’s ability to exist *when the volcano erupts and kills everyone*. Until then, it’s an active advantage. There was never a chance this would last 500 years.

If you want this to last 500 years and have any chance of detecting the next eruption that far out, you’ll need to do a great deal better. The Queen needs to not instinctively punish anyone who warns about a possible eruption, and instead let others evaluate whether or not it was a justified warning, and *reward* it if it was, while punishing failure to notice.

The Queen needs to keep track of how often alarms should go off, and get *very suspicious* if they go too many years (or generations) without a warning, and at some point assume everyone has stopped caring, fire or hang everyone involved and re-found the institute with new people.

And/or have three institute departments that don’t talk to each other but each check the volcano, and when two of them warn her but not the third, punish the third. And also probably have them do a bunch of other prediction and science tasks that keep everyone involved trained. And every so often maybe dump some strange-smelling stuff next to the volcano and have someone *claim* to have noticed something weird

and then see what reports come back on the situation. Or *something*, anything, preferably a lot of different things in unpredictable fashion.

Otherwise, I can only conclude The Volcanology Institute Is Not About Volcanos.

The Weatherman

He lives in a port town and predicts hurricanes. Hurricanes are very rare, but whenever they happen all the ships sink, so weathermen get paid very well.

If you've read your Lovecraft, you know that various sinister death cults survived the fall of Atlantis, and none are more sinister than the Cult Of The Rock. This weatherman was an adept among them and secretly communed with a rock that said "THERE WON'T BE A HURRICANE".

For many years, there was no hurricane, and he gained great renown. Other, lesser weathermen would sometimes worry about hurricanes, but he never did. The businessmen loved him because he never told them to cancel their sea voyages. The journalists loved him because he always gave a clear and confident answer to their inquiries. The politicians loved him because he brought their town fame and prosperity.

Then one month, a hurricane came. It was totally unexpected and lots of people died. The weatherman hastily said "Well, yes, sometimes there are outliers that even I can't predict, I don't think this detracts from my vast expertise and many years of success, and have you noticed some of the people criticizing me have business connections with foreign towns that probably plot our ruin?" An investigation was launched, but the businessmen and journalists and politicians all took his side, and he was exonerated and restored to his former place of honor.

Let's say The Rock is right 99.9% of the time, since that seems to be the Rule of Rocks these days, and let's say he checks it once a week. Thus, there is a hurricane roughly once every twenty years.

It sounds like worrying about hurricanes was rather expensive. Sea voyages often get cancelled. Without such worries, the town gained fame and prosperity.

This is a rather large effect. When the hurricane finally did come, yes the ships sunk and a lot of people died, but all the businessmen and journalists and politicians were cool with it. So presumably that meant that *even after the hurricane* the town was doing pretty well. Otherwise the politicians and businessmen are very much *not* going to be down with this.

There's also another possible explanation, which is that *weathermen mostly suck at predicting hurricanes*.

Have you ever seen real weathermen predict hurricanes? No. That's not a thing. Sure, they say 'it's hurricane season so there are going to be some hurricanes' and they're usually right but that very much does not count. Once a tropical storm exists and they can see it they predict how big it will get and where it is going, but that's a high-percentage play. There isn't a lot of 'three weeks from now we think there's a chance a 50% chance a hurricane is going to hit Miami.'

Whereas it sounds like other weather forecasters were essentially making those kinds of predictions often enough to seriously hurt business, and it sounds a lot like they

were mostly wrong. We have no evidence they were better than random, or that the precautions taken were net useful.

Instead of a weatherman who could plausibly have useful information, imagine an ancient tribe trying to predict a hurricane. The shaman throws bones every month, and if they land in the wrong configuration she warns of a terrible hurricane and then everyone does the ‘please don’t kill us’ prayers and ties down their stuff and then usually nothing happens because all she’s doing is throwing bones and what someone needs to do is go get a rock and carve on it “THERE WON’T BE A HURRICANE” because even if there is a hurricane it’s not like the bones were going to predict it.

Thus I applaud this Weatherman. Like in [The Phantom Tollbooth](#), he’s not actually a Weatherman, he’s more of a Whetherman. He is asked whether people should worry, and he does the right calculation and says no.

There are plenty of things like this, where the value of information of warnings is negative. Knowing you have cancer is highly useful if it can be usefully treated. If it can be wastefully treated in ways that will make you miserable and cost lots of money without doing much for your lifespan, and your doctors and family are going to push you to do that and you’ll feel guilty if you don’t, then you *really* don’t want to know. If a Covid test showing your four year old has an asymptomatic case would force lots of people to quarantine in stupid fashion because the rules are so over-the-top as to be counterproductive, maybe don’t test when you don’t have symptoms, and maybe don’t worry if the test is being done in a way that generates a bunch of false negatives. Your mother checks the weather channel, sees a 20% chance of light rain and tells you that you have to take an umbrella. One shoe bomber and we all take off our shoes at airports for a decade. And so on.

Thus, the Whetherman in question has been revealed to be doing a good job rather than a bad job. The alerts simply aren’t very specific and there are a lot of false positives, so when others warn of a hurricane there’s still only a 2% chance that it will happen at all, and *it’s not worth changing your behavior for that*. Voyages should continue, life should go on. But for various social reasons, we *can’t do that explicitly*. We can’t simply say we’re going to ignore the signs. So instead we hire someone and pay them a lot of money to give warnings, while hoping they notice that their job is bullshit and their real job is to consult the rock.

There is of course a version of this where it’s a 20% chance rather than 2%, and he got lucky for a while, and actually there were quite a lot of deaths and sunken ships and everyone is a lot worse off and he should have been fired or worse. But I am guessing we are instead in world caught in a safety trap and he did the town a favor. In those other worlds, I’m guessing that if he could actually be expected to predict hurricanes he would get to walk off into the sunset with his 20 years of fame and generous pay, but he does not get his old job back.

This raises the question of the bankers who in 2008 were caught holding a rock that said “HOUSING PRICES ALWAYS GO UP.” They had to get bailed out, which indicates the rock *wasn’t* being socially responsible, but instead one should examine the rock and it says “IF HOUSING PRICES GO DOWN YOU WOULD GET BAILED OUT.” Which is a different, perhaps smarter, rock.

This also all depends on how bad hurricanes are. A bunch of ships sinking and people dying is not great, but if the metaphorical hurricane is an existential risk like an unsafe artificial general intelligence or engineered plague that kills everyone, the calculation

looks very different. If the ‘hurricane’ is mostly harmless but everyone insists on freaking out about hurricanes, or is dangerous and causes freak outs but those freak outs do nothing useful, the wheatherman in question is a superstar, and should be paid very well indeed.

Rock is Strong

Scott worries that experts who are charged with spotting rare events will, instead of doing the real work, end up relying on the 99.9% accurate heuristic, and that this will be bad. Here’s his explanation.

Maybe this is because the experts are stupid and lazy. Or maybe it’s social pressure: failure because you didn’t follow a well-known heuristic that even a rock can get right is more humiliating than failure because you didn’t predict a subtle phenomenon that nobody else predicted either. Or maybe it’s because false positives are more common (albeit less important) than false negatives, and so over any “reasonable” timescale the people who never give false positives look more accurate and get selected for.

You say ‘stupid and lazy’ and I say ‘respond to incentives’ and ‘cannot actually do much better than the heuristic if they aren’t allowed to give probabilistic answers.’ Or ‘the heuristic isn’t that easy to implement, you try it.’ You say ‘didn’t follow a well-known heuristic even a rock could follow’ and I say ‘didn’t do the job they were actually hired to do.’ You say ‘look more accurate and get selected for’ and I say ‘evaluate on results rather than process.’

(Also, Scott says ‘99.9% accurate’ and I say ‘I very much doubt that.’)

More than all that, Scott worries that And That’s Terrible.

This is bad for several reasons.

First, because it means everyone is wasting their time and money having experts at all.

But second, because it builds false confidence. Maybe the heuristic produces a prior of 99.9% that the thing won’t happen in general. But then you consult a bunch of experts, who all claim they have *additional* evidence that the thing won’t happen, and you raise your probability to 99.999%. But actually the experts were just using the same heuristic you were, and you should have stayed at 99.9%. False consensus via [information cascade!](#)

This new invention won’t change everything. This emerging disease won’t become a global pandemic. This conspiracy theory is dumb. This outsider hasn’t disproven the experts. This new drug won’t work. This dark horse candidate won’t win the election. This potential threat won’t destroy the world.

All these things are *almost* always true. But Heuristics That Almost Always Work tempt us to be more certain than we should of each.

You say ‘wasting their time and money on experts,’ I say ‘giving people peace of mind and some combination of social and legal cover to do what they want to do anyway without blame’ and ‘the heuristic has non-obvious detail the experts are evaluating’ and ‘think of all the money we save by having the experts work so quickly.’

You say ‘building false confidence’ and I say that’s mostly on us not the ‘experts.’

Once I notice that there is a heuristic this accurate, I should also notice that the experts correlate highly with such accurate heuristics, and that they are probably not that much more accurate than the heuristics, so I mostly should not update much. At this kind of extreme it’s not even an issue. If an ‘expert’ tells me X is false but the basic heuristics say X is 99.9% false, what’s my new probability? If the expert told me this without my asking directly, probably *lower* than 99.9%, because the fact that they felt the need to tell me is more important than their opinion here.

A more scary situation is if the heuristic clocks in at 90% and the experts copy it, and now I’m going to plausibly end up substantially higher than 90%, but I definitely noticed both so I should notice that there’s some causation there in terms of the heuristic being known to the experts. But yes, I should totally update a substantial amount, because the experts are at least confirming that I have correctly assessed the reference classes involved. Couldn’t be sure about that.

What the heuristic is *actually for* here is to push the experts towards it, because those who don’t consider the heuristics end up not aligning with them enough, and they need to update and keep in mind that one should not go against it lightly, instead considering other possibilities first. And also because forgetting or neglecting the heuristic will waste a lot of valuable time. Dr. House often reminds us that it’s never lupus, but one time it *is* lupus, and he does eventually figure this out in the end.

There’s also the strong possibility, which Scott does not consider but that seems prominent in most of Scott’s examples, that the expert is actually being paid mostly to use the heuristic so it can be seen as coming from an expert and people can be assured there is a ‘human in the loop’ who is checking for the obvious exceptions. Such jobs are mostly or entirely bullshit jobs.

Often you don’t care about knowing, but you don’t want others to know you don’t care about knowing. If they knew you didn’t care (and didn’t care about others learning you didn’t care and so on) then that could mean you don’t care, didn’t take precautions or are otherwise blameworthy if things go wrong. Other times, your ignorance can be exploited, and they can rob the place.

Mostly this is all a case of People Respond to Incentives. I notice that many of the plays here don’t require that the heuristic be accurate, certainly not at anything like the 99.9% level. They only require that payoffs not properly correspond to outcomes. I don’t need an accurate heuristic to know that if warning about the volcano (or cancer, or housing prices, or a robbery, or anything else) only gets me in trouble, it doesn’t matter what I suspect with what probability, there’s a right answer to ‘what should I say in this spot?’ and that is what you are probably going to hear. And over time, people figure out that sort of thing quite well.

I also notice Scott’s examples are full of *vast overconfidence*. Who are these people who have 99.9% accurate heuristics while constantly making important mistakes? That’s quite a neat trick.

That leads into the comparison to black swans and tail risks. Sometimes the issue here is the assumption of tail risk and the possibility of a black swan. When the volcano erupts, the hurricane comes or the market crashes, or the invention changes the world, the impact of that is huge, so always saying ‘no’ is almost always right and also a bad trade that gets super expensive, or is effectively counting on a bailout where the person in question will not end up having to pay the bill that comes due.

I agree with Scott that this is a *subset* of what is happening here. Even in the cases where there is a tail risk that people are laying off on others, that problem alone is only some of what has gone wrong, and would not alone cause the same level of pickle. In other cases it is entirely distinct from the issue. Giving everyone a mysterious impossibly high 99.9% accuracy made this seem more central than it was.

When the problem is the failure of such heuristics to price in the costs of tail risks, that can be a rather large mistake, but also sometimes the tail risk is being overpriced rather than underpriced. In those cases where there are real and contextually large tail risks, such as artificial general intelligence or more typically with a trading algorithm, there is a big problem. Other times, the big problem is the perception of a potential big problem, and the actual problem is small.

I also wanted to ensure I pointed to what I see as several conceptual errors on Scott's part, that I worry are indicative of fundamental model errors - the idea that simple-sounding heuristics are actually simple to execute in practice, the assumption that jobs are not bullshit and that people want those doing those jobs to be accurate rather than do something else, especially the symbolic and signaling values involved, and the failure to think about the value of information and the cost of getting that information. There is a kind of 'naïve mistake theory' here that is sufficiently naïve that it does not seem like a mistake.

I have split off the postscript to this, entitled [Paper is True](#), into its own post.

Why Doesn't Healthcare Improve Health?

This is a linkpost for <https://www.epistem.ink/p/why-doesnt-healthcare-improve-health?r=c5fr0>

It's almost a non-sequitur to ask whether or not a giant industry, with a well-defined goal, churning double-digits of GDP, actually works.

If pressured into an experiment we could easily design one that shows airplanes deliver people to destinations over 3000km away faster than trains (boarding time and travel to airport included). We could similarly prove they help people travel more, and more money spent on airplanes means more travel. Take a random group of 1000 people, give half 5000\$ airline credits for two years, and we'll certainly find that, on average they end up visiting more far-away places than the control.

On the other hand, this is not entirely unheard of.

A big and controversial example is the church. Where we've pretty conclusively proven that building huge cathedrals, converting "savage natives", stomping out heresy and prayer have little results upon any form of well-being. But even here we must admit that, if the apparatus of science is unleashed upon religion, we will find interesting things such as people in religious communities drinking less, having longer-lasting marriages, and living longer with higher self-reported life satisfaction. We can argue about confounders and about the causal mechanism being replicable with things other than religion. But I think that, to most scientifically minded people, the lack of evidence is sufficient to *not* motivate them into becoming religious.

However, I keep being surprised by intelligent and scientifically minded people outright refusing to hear the evidence **against** medicine being useful. That's not to say evidence against any particular therapy. There are many things that obviously work, such as emergency intervention for losing a limb in a car crash, where "almost certain death" is the clear outcome. There are many things that obviously work from a statistical perspective, such as vaccination against diseases such as covid, hep A & B, polio, HPV, yellow fever, rabies, and the like. There are also many things that seem to work, with good mechanistic evidence and promising studies, such as PRP for various forms of soft-tissue injury, or prophylactic therapy for HIV.

But this therapy-specific evidence is only more damning when viewed in the context of broader findings against medicine. Robin Hanson [has written on this ad-nauseam](#), so I will let him do most of the talking.

1974 to 1982 the US government spent \$50 million to randomly assign 7700 people in six US cities to three to five years each of either free or not free medicine, provided by the same set of doctors. ... people randomly given free medicine in the late 1970s consumed 30-40% more medical services, paid one more "restricted activity day" per year to deal with the medical system, but were not noticeably healthier! ([More](#), see [also](#))

and

Oregon assigned a limited number of available Medicaid slots by lottery. ... 8,704 (~30%) [very sick and poor US adults] were enrolled in Medicaid medical insurance. ... at most see two years worth of data. ... had substantially and significantly better self-reported health. ... over two thirds of the health gains ... appeared on the very first survey, done before lottery winners got additional medical treatment. (More)

No statistically significant effect on measures of blood pressure, cholesterol, or blood sugar. ... did not reduce the predicted risk of a cardiovascular event within ten years and did not significantly change the probability that a person was a smoker or obese. ... it reduced observed rates of depression by 30 percent. ([More](#))

and

This study ... is amongst the largest health insurance experiments ever conducted ... in Karnataka, which spans south to central India. The sample included 10,879 households (comprising 52,292 members) in 435 villages. Sample households were above the poverty line ... and lacked other [hospital] insurance. ... randomized to one of 4 treatments: free RSBY [= govt hospital] insurance, the opportunity to buy RSBY insurance, the opportunity to buy plus an unconditional cash transfer equal to the RSBY premium, and no intervention. ...intervention lasted from May 2015 to August 2018. ...

Opportunity to purchase insurance led to 59.91% uptake and access to free insurance to 78.71% uptake. ... Across a range of health measures, we estimate no significant impacts on health. ... We conducted a baseline survey involving multiple members of each household 18 months before the intervention. We measured outcomes two times, at 18 months and at 3.5 years post intervention. ... only 3 (0.46% of all estimated coefficients concerning health outcomes) were significant after multiple-testing adjustments. We cannot reject the hypothesis that the distribution of p-values from these estimates is consistent with no differences (P=0.31). ([more](#))

But I think Hanson overlooks one of [the best and funniest studies ever conducted](#), based on dutch health insurance data.

Data from 1913 conventional GPs were compared with data from 79 GPs with additional CAM training in acupuncture (25), homeopathy (28), and anthroposophic medicine (26). Results Patients whose GP has additional CAM training have 0–30% lower healthcare costs and mortality rates, depending on age groups and type of CAM. The lower costs result from fewer hospital stays and fewer prescription drugs

The funny bit here is that not only are all 3 types of alternative medicine trained GPs better, both when taken in aggregate and individually. But the best outcomes seem to come out of homeopathy, which is as perfect of a placebo arm as one can get.

This is not to say that *all* evidence points towards medicine having no effect whatsoever on health-related biomarkers, quality of life, or mortality. It's just that most studies point towards this being the case, and the few that don't, aren't finding very significant effects.

This is surprising, given that we **know** a bunch of interventions clearly work, really well.

Regardless, I started with the problem of *why* well-educated people seem to ignore this evidence and chose to use, pay for, subsidize, pay others to, and lobby for subsidization of healthcare services.

Surely if there was evidence pointing towards more deaths, longer travel times, and increased costs of going by plane instead of train on most or all routes in Europe or the US, most of these people would start using trains almost exclusively. Why is the same argument not clicking [here](#)?

I think the issue is, at least in part, with a lack of understanding as to **why** this is happening. What parts of healthcare are broken. This in itself is a problem I can't address because few studies are done trying to address this since the raison d'etre for these studies is being cognitively suppressed by most people in a position to run them.

Still, I hope that the evidence against medicine might be easier to swallow if people at least had some hypothesis for why it might not work, and which bits of it might work. So here are three such hypotheses, which in part make my own working model of healthcare.

i - Diagnosis Is Broken

It might be that this standard of rigor is *not* being applied during a normal diagnosis procedure. The evidence here is hard to assess since there's no meta-analysis of the subject as a whole.

When most people run studies on interventions they are very careful to pick patients that actually have the condition being studied. There are exclusion criteria, both during the trial and during the analysis of the data. Doctors are instructed very carefully when to prescribe the new intervention.

But let me take the Dutch insurance trial as a potential example here. A comment I found [here](#) on the Dutch medical system is:

“go home, take some Tylenol and come back if you don’t feel better” is actually quite an effective strategy in this GP-as-gatekeeper model. Most of your patients feel better and don’t come back, as you couldn’t have done anything for them anyway. This keeps costs down and keeps the emergency room just for actual emergencies.

I think most people agree that, if someone comes into a GP complaining of head/back/stomach pain, just prescribing some acetaminophen as a placebo, instead of recommending investigations (which induce anxiety, cost money, and will find nothing) or prescribing opioids, is preferable.

... Except that most people forget acetaminophen is actually [quite a dangerous drug](#).

Acetaminophen overdose is the leading cause for calls to Poison Control Centers (>100,000/year) and accounts for more than 56,000 emergency room visits, 2,600 hospitalizations, and an estimated **458 deaths** due to acute liver failure each year.

This is a first-order effect, ignoring the second-order effects of acetaminophen, which essentially stops inflammation processes inside the body that are critical for stopping

pathogens and signaling cells responsible for healing tissue injuries.

So it might well be that a GP prescribing an actual placebo (e.g. a homeopathic pill) would have their patients fare much better. Or that someone never going to a doctor for backache and just “ignoring it” would fare much better. Even in the case where the GP recommends what is seen as a “completely safe treatment”.

Similarly, many such small interventions for issues that require non might be accumulating to cause long-term health issues in the long run. Be it overprescribing SSRIs for mild cases of depression, thus stopping people from trying to solve their actual issues, or fixing a minor & asymptomatic cavity “just in case”, thus damaging the structural integrity of teeth and inflicting infection vectors upon the patient for no reason.

I've personally had dentists recommend to me that I should totally shatter my mandible, do open surgery to transplant some bone from another part of my body, then wait a few months for it to heal back into a shape that might “improve my bite”.

I've also had really good doctors recommend surgeries for [conditions that don't improve with surgery](#), and where surgery is known to cause long-term health deterioration.

Not to mention I've had doctors recommend me “standard” procedures such as using vitamin A creams of acne, a product which, in the case of someone with high serum vitamin A like myself, would lead to liver toxicity without any evidence of improving health outcomes. Or use sunscreen, in spite of dozens of studies showing mostly no links, or even harmful links, [between sunscreen use and various skin cancers](#).

Add to this common-sense “dumb” things such as doctors recommending powerful bleaching skin “treatments” and invasive investigations (contrast MRIs, arthroscopic investigations, endoscopy, colonoscopy, biopsies, radiology).

The problem with most “minor” treatments that get commonly recommended is that they are not dangerous enough to show serious side effects, and there are no incentives to run in-depth controlled studies to look for minor problems. There might be hundreds of other “minor” doctor recommendations that cause cumulative harm, which in of themselves are minor, but pile up onto someone that visits doctors dozens of times a year.

Now add onto that the fact that similar misunderstanding of evidence and misdiagnosis by doctors increases the chance of *not* being recommended actual life-improving treatments that have shown their worth in clinical trials where doctors were explicitly trained to use them.

Even if we have a suite of interventions that have mild effects when prescribed correctly, they also have side effects. If the prescription methodology of doctors is unable to select interventions correctly, we might have an effect that showcases minor improvements in clinical trials but leads to harm in “real” usage.

ii - Randomized Placebo-Controlled Trials Don't Work

Controlled trials are a great tool to weed out ineffective medicine. Most doctors will use success in RCTs with large sample sizes plus FDA/EMA approval as reason enough to prescribe something. Even the most cautious doctors will probably give in once conclusively shown that something works by an impartial meta-analysis run by specialists on multiple studies.

If you agree with this methodology, great. I once again invite you to go to your local shaman and buy some [homeopathic medicine](#) (use scihub to open the study). It has better evidence than most drugs, in summary, it shows:

The combined odds ratio for the 89 studies entered into the main meta-analysis was 2.45 (95% CI 2.05, 2.93) in favour of homeopathy. The odds ratio for the 26 good-quality studies was 1.66 (1.33, 2.08), and that corrected for publication bias was 1.78 (1.03, 3.10). Four studies on the effects of a single remedy on seasonal allergies had a pooled odds ratio for ocular symptoms at 4 weeks of 2.03 (1.51, 2.74). Five studies on postoperative ileus had a pooled mean effect-size-difference of -0.22 standard deviations (95% CI -0.36, -0.09) for flatus, and -0.18 SDs (-0.33, -0.03) for stool (both $p < 0.05$).

Obviously, we'll just go ahead and dismiss homeopathy based on the underlying mechanism and the fact that these trials show minor results and are run by motivated institutions which are prone to slightly altering numbers, not publishing any negative results, and doing as much statistical manipulation as possible without revealing the actual data unless required.

... But the exact same thing can be stated about many therapies, ranging from statins to aducanumab. Yet doctors seem perfectly happy to prescribe a lot of low-impact drugs.

Even worst, unlike placebo pills, low-impact drugs might actually have hidden side effects.

This gets even worst with interventions against malignancies such as cancer, which get a much easier pass when it comes to [what's considered efficacious](#). And "best practices" might encourage drugs with evidence for increased mortality.

That is not to say RCTs are a bad model. If effects are stunning enough, then an RCT should count as the definitive proof that a drug works. But if the results are minor we might be misled by research bias. This is a broader problem with modern science, which often ignores the importance of effect magnitude versus direction.

iii - Hospitals Are Dangerous

Going to the hospital is fairly common for conditions that are life-impairing but likely not fatal. Not to mention that people often get surgeries and engage in multi-day ICU stays for treatment.

Hospitals are primarily dangerous to anyone, even someone going for a simple consultation or to get an MRI, due to the presence of many sick people, some

infectious, and due to [the abundance of drug-resistant pathogens](#), which are mainly absent from other environments.

Once we get into receiving invasive treatment this gets worst. We have to keep in mind that most surgeries vs conservative trials are run at good hospitals, the kind that have doctors interested in running trials, and we might expect that doctors knowing they are participating in trials take extra care not to do things like... accidentally kill patients.

But accidental deaths are a huge concern when doing any sort of surgery, even elective ones. [Stealing from Hanson](#) again:

In 1999, the Institute of Medicine published the famous “To Err Is Human” report, ... reporting that up to 98,000 people a year die because of mistakes in hospitals. The number was initially disputed, but is now widely accepted by doctors and hospital officials — and quoted ubiquitously in the media. In 2010, the Office of Inspector General for Health and Human Services said that bad hospital care contributed to the deaths of 180,000 patients in Medicare alone in a given year.

Now comes a study in the current issue of the Journal of Patient Safety that says the numbers may be much higher — between 210,000 and 440,000 patients each year who go to the hospital for care suffer some type of preventable harm that contributes to their death, the study says. That would make medical errors the third-leading cause of death in America, behind heart disease, which is the first, and cancer, which is second.

James based his estimates on the findings of four recent studies that identified preventable harm suffered by patients – known as “adverse events” in the medical vernacular – using a screening method called the Global Trigger Tool, which guides reviewers through medical records, searching for signs of infection, injury or error. Medical records flagged during the initial screening are reviewed by a doctor, who determines the extent of the harm.

In the four studies, which examined records of more than 4,200 patients hospitalized between 2002 and 2008, researchers found serious adverse events in as many as 21 percent of cases reviewed and rates of lethal adverse events as high as 1.4 percent of cases.

By combining the findings and extrapolating across 34 million hospitalizations in 2007, James concluded that preventable errors contribute to the deaths of 210,000 hospital patients annually.

That is the baseline. The actual number more than doubles, James reasoned, because the trigger tool doesn’t catch errors in which treatment should have been provided but wasn’t, because it’s known that medical records are missing some evidence of harm, and because diagnostic errors aren’t captured. An estimate of 440,000 deaths from care in hospitals “is roughly one-sixth of all deaths that occur in the United States each year.” ([more](#); [source](#))

That seems rather horrible, but again, it doesn’t cover long-term damage from hospital stays such as non-fatal infections.

Finally, people opting out of medicine aren’t taking a vacation to visit a sunless brutalist building with sad people and agitation when they feel sick. They might

instead call in sick and spend time playing board games with family, start eating better, go to the beach, or just sleep more.

iv - What I Do

I think the evidence against the efficacy of healthcare is pretty damning, and it scares me.

I don't believe it boils down to just the above reasons and I don't believe I have enough evidence to say they are true.

The one thing I refuse to do is close my eyes to the evidence of medicine not working. This is especially important since **some** parts of medicine obviously work, and the upside here is not just saving money, it's significantly increasing my healthspan and lifespan by being very selective about the medicine I use.

My current protocol is something like this:

Don't register problems as "medical" problems unless they are really bad or unless I have the mental time to investigate a solution. If my head hurts, my head hurts. If my head hurts every day for 2 weeks, that's a medical issue. But I will try to avoid using medicine when I can rely on homeostasis.

Don't heed **any** advice from doctors about elective treatment, look at the direct evidence myself. Doctors can be hypothesis generators and they can help conceptualize problems. They can turn "indistinct pain here" into "MRI shows inflammation in foobar muscle", this is hugely useful for actually being able to look at the relevant evidence. Though one bit that shouldn't be forgotten is that the diagnosis itself might be wrong, and looking at the error rates on the diagnosis you're getting is an important thing to do **before** taking it as a data point. You have access to all the information a doctor has, you might be much worse at parsing through it, but you can afford to spend 100 times the amount of time a doctor would on your case.

Refuse all surgical interventions that aren't life-saving outside of extraordinary circumstances. Obviously don't refuse life-saving surgery after a car crash.

Avoid therapies that don't have tremendously large magnitudes associated with their effect direction **unless** you are directly monitoring some biomarkers and have large volumes of evidence and mechanistic reasons showing no side effects.

Avoid consuming healthcare without significant time to reason about it. It takes time to go directly into the evidence instead of heeding doctor advice, so if you have dozens of ongoing conditions you won't be able to handle all of them. The only solution is to try to prioritize, focus on the important ones, and forget the milder issues until you've solved the important ones... and hey, maybe they'll heal on their own.

This is still not an ideal solution, there is vague semantics here for which I don't have strict definitions.

If I'm mixing a multi-vegetable and algae powder with my morning yogurt... is that a medical intervention? A supplement? Just a way of consuming food?

If I go get a hot stones massage is that “medicine”? What about a sports massage at a qualified therapist?

Is doing some yoga for back pain a medical intervention? What about following a routine prescribed by a PT?

I don't know.

The danger of not trusting medicine enough is starting to trust quackery too much. The way I solve this is by self-experimenting and finding broad-spectrum high-impact interventions I can always use.

But what will I do when I have a serious issue, no time or ability to interpret evidence (or no evidence to interpret), and my catch-all interventions fail? Do I just trust a surgeon? Do I go to a spiritual healer? Do I just ignore things like black liquid sipping out of multiple orifices?

I don't know, my system has edge-cases which I'm sure will make it broken for people that aren't lucky enough to be in their 20s. This is why my principal focus is still on trying anything and everything I can to monitor and delay aging.

The only thing I will say is that we have a long history where truth-seeking in the face of uncomfortable circumstances seems to work out. So we shouldn't make ourselves immune to evidence just because the things it says will cause discomfort. We should figure out the extent to which we should trust it and integrate it into our lives, then hope for the best.

... or who knows, maybe we should just trust the evidence fully and abandon our life to [join a California doomsday cult](#). I hear the epidemiological studies find it quite promising.

Naturalism

According to me, this sequence has been pretty darn abstract.

That was kind of on purpose. It's the opposite of what I like to do, of what I think I'm good at. I much prefer to engage with an actual specific *thing*, and to share the details of my experience as I go. This big picture stuff is really not my jam.

But I've been trying to paint a really big picture anyway, to describe an entire perspective on investigation, and rationality, and maybe life. I hope it's been much easier to read than it was for me to write. And I hope that if, at some future point, I dive into the little details of particular exercises and techniques, you'll be able to contextualize them as more than just trinkets, or rituals that are tedious to little purpose.

But I'm so tired of it. I'm exhausted by all this abstraction. I want to touch the ground. I want to show you what it actually looks like to live a life full of patient and direct observation.

I can tell you that there's a magnifying glass in my pocket, which I use regularly. I can tell you that I turned the soles of my bare feet toward the sky last week, so that I could feel the snow falling on them. I can tell you that when I put "it seems to me" at the front of so many of my sentences, it's not false humility, or insecurity, or a verbal tic. (It's a deliberate reflection on the distance between what exists in reality, and the constellations I've sketched on my map.)

I can tell you dozens of facts like these, about my experience of myself and of the world. Hundreds. But none of those means much. Not on its own.

The problem is, this whole thing is founded on patience, which is difficult to demonstrate in an essay. It's hard to show you all at once the myriad ways a thousand tiny moments add up to one big thing that matters.

Still, they do add up to something.

What they add up to is that I am a naturalist. I was raised to be a naturalist, and it worked. I was raised to be someone who *yearns* to know the territory through patient and direct observation. My childhood memories are full of mushroom hunting, finding newts under logs, following game trails, reading the geological histories in the rock layers whenever we traveled, and sketching the paths of Jupiter's moons with a red flashlight beside my telescope.



My upbringing emphasized that the world is an infinity of wonders; unfathomably many in a single handful of dirt. It taught me that knowledge is power. It taught me that although school and books and the edifice of scientific inquiry can help you orient and make sense of your observations, there is exactly *one* key in the whole universe that can unlock the power of knowledge—and that key is your eagerness to go out into the world, day after day, and look with your own eyes at what is in front of you.

There isn't space in a concluding essay to properly describe the habits comprising this way of life, or their result. But if I've communicated even half of what I hoped to in this sequence, you may now be in a good position to find out for yourself.

Think of some problem you have, something you want to get a better handle on or otherwise figure out. Maybe it's something to do with your career path, a place where you're stuck in your research, or the way you spend time with your kids. Anything where you're yearning for deeper, more masterful knowledge than you have right now.

(There can be a lot of inertia in the flow from paragraph to paragraph. Here is a place to pause. Even if you're not up for a thought experiment right now, I request that you count to twelve before reading on, just in case something comes to you by accident as soon as I've stopped shouting words into your head.)

Now imagine that there's no internet, and not a single expert available to advise you. Your only books are the ones you write. Your only resources are your body, your mind, and the world itself.

If I wanted to know morel mushrooms, I would look for them beneath an old hardwood tree in a Midwestern forest in spring. I'd go there right about the time the mayapples are in bloom. I

would look at the ground, in damp places where the autumn leaves have partially decomposed. That is the natural habitat of morels.

What is the natural habitat of the thing that interests you? Where could you go to observe it directly? How could you invite it to impinge on your experience? And what, if anything, is in the way of you being *open* to it when it does?

If you're not sure of its natural habitat, then what's your best guess, and how could you tell when you're getting warmer? What might tip you off that some tendril of the thing's reality has just brushed your mind? How might you recognize if *now* is the time to pay attention, and to make a new guess about where to look next?

And what could you do to observe it over time, to see beyond your very first impression? What little habits might you adopt, like an athlete who always takes the stairs, to ensure that you make frequent contact with this patch of territory in daily life? How might you record your observations, and notice patterns that aren't apparent in any single moment?

If you wanted to increase your contact with the world, what is the very first thing you would change?

Knowing the territory takes patient and direct observation.

This is what I mean by "naturalism".

Whence the sexes?

Note: I've edited the post to swap explanations #5 and #6, and added new counterarguments against the importance of #5 (self-fertilisation).

There are many explanations of the evolutionary value of sex in terms of gene exchange ([I particularly like this one](#)). But these don't explain the evolutionary value of having sexes: of the differentiation between males and females. A species of [hermaphrodites](#) would get all the genetic benefits of sex, but without the massive cost of half its population being unable to bear offspring. On average, each individual could have twice as many offspring, unless other problems arose. And indeed, most plants are hermaphroditic - but only a few animals. So why aren't most animals hermaphrodites? A quick search doesn't turn up any widely accepted answer, so I've brainstormed a few possibilities. I may well be missing something obvious; if so, let me know.

1. **Resource cost of being hermaphroditic.** If there's a strong [division of labour](#) between males and females, then maybe it's harder for hermaphrodites to gather enough resources to support offspring. But in many species the males contribute little in terms of resources - e.g. in orang-utans, who are very solitary.
2. **Developmental or metabolic costs of being hermaphroditic.** Maybe it's just a very expensive adaptation. But this seems unlikely to be the main factor - the relevant baseline is the existence of males at all, which is a huge energy cost.
3. **Difficulty of evolving hermaphroditism.** Maybe this is just hard for evolution to find? But non-reproductive hermaphroditism seems like it arises via mutations pretty frequently, so I'd be surprised if reproductive hermaphroditism were unachievable by evolution. And in fact there are a few hermaphroditic species - so why haven't they spread much more widely?
4. **Difficulty of fixating hermaphroditism.** A hermaphrodite in a species without many hermaphrodites is likely not as attractive to females as most males are, nor as fertile as most females are. So maybe, even after arising, the trait will be selected against. But on timeframes where sexual desires can themselves evolve, all else equal we should expect stabilising sexual selection towards the best combination of fertility and attractiveness. E.g. it would be undesirable to be impregnated by overly masculine conspecifics, because the resulting offspring would be less fertile themselves. So this adds a bit more difficulty to reaching the hermaphroditic equilibrium, but doesn't answer the core question of why that's a less fit equilibrium.
5. **Self-fertilisation.** I remember reading a while back that self-fertilisation has a strong short-term advantage, despite losing out on the long-term benefits of sex ([this old paper](#) has a section on "selection in favour of self-fertilisation", although I haven't read it in detail). So maybe the answer is that hermaphrodites end up evolving ways to self-fertilise, which is harmful in the long run, and so group selection prevents the trait from becoming too widespread. This effect [plausibly occurs in plants](#), but not strongly enough to prevent most of them from being hermaphroditic. And presumably it's significantly harder for plants to prevent self-fertilisation (since their pollen spreads widely) than it is for animals.

I'm open-minded to the possibility that a combination of these explanations is responsible. But none of them seems particularly strong to me; so I'm guessing that the biggest effect comes from:

6. Physical dominance. Maybe animal competition to impregnate fertile conspecifics is grounded in physical power, so that dominant males could just prevent hermaphrodites (who invest less in muscle and brawn) from having sex. In some sense this is a variant of the first possibility: the comparative advantage of muscle is just so strong that the best solution is to have a division of labour. But it focuses not on problems posed by the environment, but rather on problems posed by one's own species. If true, it feels a bit sad: that there could be a much better solution if it weren't for the threat of physical force. But it does seem pretty plausible to me - especially because hermaphroditism is much more common in plants, which can't use the strategy of physical dominance.

- The main argument against it is that males in some species don't compete via shows of force - e.g. birds which sing to attract mates. But birds are unusual in other ways too - e.g. over 90% of bird species are monogamous (as compared with less than 5% of mammals), which makes it more plausible that the "strong division of labour" hypothesis explains their sexual differentiation. So I'd be interested in pointers to any literature on how many non-monogamous species lack physical male competition.
- Also, in [some species](#) several different mating strategies remain in equilibrium (including strategies which involve surreptitious mating unnoticed by a dominant male). So even if physical dominance is the best strategy, is it really so much better that it can crowd out all the others?

Getting more clarity on this topic isn't a priority for me, but I do think of the question as one small data point that might help ground big-picture abstractions about competition and cooperation (in a comparable way to how knowledge of how insect colonies works provides an interesting metaphor for thinking about society).

[Intro to brain-like-AGI safety] 3. Two subsystems: Learning & Steering

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

3.1 Post summary / Table of contents

Part of the [“Intro to brain-like-AGI safety” post series](#).

In [the previous post](#) I defined the notion of “learning from scratch” algorithms—a broad category that includes, among other things, any randomly-initialized machine learning algorithm (no matter how complicated), and any memory system that starts out empty. I then proposed a division of the brain into two parts based on whether or not they learn from scratch. Now I’m giving them names:

The **Learning Subsystem** is the 96% of the brain that “learns from scratch”—basically the telencephalon and cerebellum.

The **Steering Subsystem** is the 4% of the brain that *doesn’t* “learn from scratch”—basically the hypothalamus and brainstem.

(See [previous post](#) for a more detailed anatomical breakdown.)

This post will be a discussion of this two-subsystems picture in general, and of the Steering Subsystem in particular.

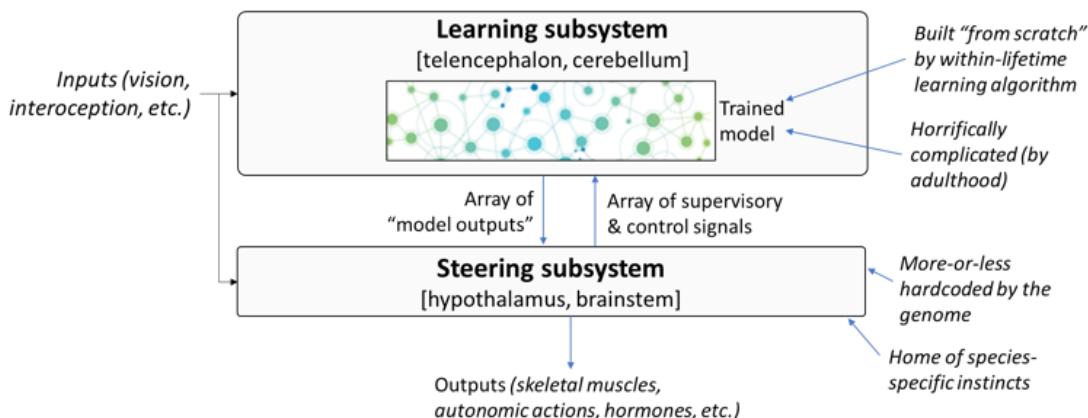
- In Section 3.2, I’ll talk about the big picture of what these subsystems do and how they interact. As an example, I’ll explain why each subsystem needs its own sensory-processing circuitry—for example, why visual inputs get processed by *both* the visual cortex in the Learning Subsystem, *and* the superior colliculus in the Steering Subsystem.
- In Section 3.3, I’ll acknowledge that this two-subsystem picture has some echoes of the discredited “triune brain theory”. But I’ll argue that the various problems with triune brain theory do not apply to my two-subsystem picture.
- In Section 3.4, I’ll discuss three categories of ingredients that could go into a Steering Subsystem:
 - Category A: Things that are plausibly essential for general intelligence (e.g. an innate drive for curiosity),
 - Category B: Everything else in the human steering subsystem (e.g. an innate drive to be kind to your friends),
 - Category C: Any other possibility that an AGI programmer might dream up, even if it’s radically different from anything in humans or animals (e.g. an innate drive to correctly predict stock prices).
- In Section 3.5, I’ll relate those categories to how I expect people to build brain-like AGIs, arguing that “*brain-like AGIs with radically non-human (and dangerous) motivations*” is not an oxymoron; rather, it’s the default expected outcome, unless we work to prevent it.
- In Section 3.6, I’ll discuss the fact that Jeff Hawkins has a two-subsystems perspective similar to mine, yet argues *against* AGI catastrophic accidents being a risk. I’ll say where I think he goes wrong.
- Sections 3.7 and 3.8 will be the final two parts of my “timelines to brain-like AGI” discussion. The first part was [Section 2.8 in the previous post](#), where I argued that reverse-engineering the Learning Subsystem (at least well enough to enable brain-like AGI) is something that could plausibly happen soon, like within the next decade or two, although it could also take longer. Here, I’ll complete that story by arguing that this same thing is true of reverse-engineering the Steering Subsystem (at least well enough to enable brain-like AGI), and of getting the algorithms cleaned up and scaled up, running model trainings, and so on.

- Section 3.9 is a quick non-technical discussion on the wildly divergent attitudes that different people take towards the timeline to AGI, even when they agree on the probabilities. For example, you can have two people agree that the odds are 3:1 against having AGI by 2042, but one might emphasize how low that probability is (“You see? AGI probably isn’t going to arrive for *decades*”), while the other might emphasize how *high* that probability is. I’ll talk a bit about the factors that can underlie those attitudes.

3.2 Big picture

In [the last post](#), I claimed that 96% of the brain by volume—roughly the telencephalon (neocortex, hippocampus, amygdala, most of the basal ganglia, and a few other things) and cerebellum—“learns from scratch”, in the sense that early in life its outputs are all random garbage, but over time they become extremely helpful thanks to within-lifetime learning. (More details and caveats in [the previous post](#).) I’m now calling this part of the brain the **Learning Subsystem**.

The rest of the brain—mainly the brainstem and hypothalamus—I’m calling the **Steering Subsystem**.



How are we supposed to think about these?

Let’s start with the Learning Subsystem. As discussed in [the last post](#), this subsystem has some interconnected, innate learning algorithms, with innate neural architectures and innate hyperparameters. It also has *lots* (as in billions or trillions) of adjustable parameters of some sort (usually assumed to be synapse strength, but this is controversial and I won’t get into it), and the values of these parameters start out random. The Learning Subsystem’s algorithms thus emit random unhelpful-for-the-organism outputs at first—for example, perhaps they cause the organism to twitch. But over time, various supervisory signals and corresponding update rules sculpt the values of the system’s adjustable parameters, tailoring them within the animal’s lifetime to do tricky biologically-adaptive things.

Next up: the Steering Subsystem. How do we think intuitively about that one?

First off, imagine a repository with lots of species-specific instincts and behaviors, all hardcoded in the genome:

- “In order to vomit, contract muscles A,B,C, and release hormones D,E,F.”
- “If sensory inputs satisfy the *thus-and-such* heuristics, then I am probably eating something healthy and energy-dense; this is good and I should react by issuing signals G,H,I.”
- “If sensory inputs satisfy the *thus-and-such* heuristics, then I am probably leaning over a precipice; this is bad and I should react by issuing signals J,K,L.”
- “When I’m cold, get goosebumps.”

- “When I’m under-nourished, do the following tasks: (1) emit a hunger sensation, (2) start rewarding the neocortex for getting food, (3) reduce fertility and growth, (4) reduce pain sensitivity, etc.” ([ref](#)).

An especially-important task of the Steering Subsystem is sending supervisory and control signals to the Learning Subsystem. Hence the name: the Steering Subsystem *steers* the learning algorithms to do adaptive things.

For example: How is it that a *human* neocortex learns to do adaptive-for-a-human things, while a *squirrel* neocortex learns to do adaptive-for-a-squirrel things, if they’re both vaguely-similar learning-from-scratch algorithms?

The main part of the answer, I claim, is that the learning algorithms get “steered” differently in the two cases. An especially important aspect here is the “reward” signal for reinforcement learning. You can imagine that the human brainstem sends up a “reward” for achieving high social status, whereas the squirrel brainstem sends up a “reward” for burying nuts in the fall. (This is oversimplified; I’ll be elaborating on this story as we go.)

By the same token, in ML, the *same* learning algorithm can get really good at playing chess (given a certain reward signal and sensory data) *or* can get really good at playing Go (given a *different* reward signal and sensory data).

To be clear, despite the name, “steering” the Learning Subsystem is but one task of the Steering Subsystem. The Steering Subsystem can also just up and do things, all by itself, without any involvement from the Learning Subsystem! This is a good plan if doing those things is important right from birth, or if messing them up even once is fatal. An example I mentioned in [the last post](#) is that mice apparently have a [brainstem bird-detecting circuit wired directly to a brainstem running-away circuit](#).

An important dynamic to keep in mind is that the brain’s Steering Subsystem cannot directly access our common-sense understanding of the world. For example, the Steering Subsystem can implement reactions like “when eating, manufacture digestive enzymes”. But as soon as we start talking about the abstract concepts that we use to navigate the world—grades, debt, popularity, soy sauce, and so on—we have to assume that the Steering Subsystem has no idea what any of things are, unless we can come up with some story for how it found out. And sometimes there *is* such a story! We’ll see a lot of those kinds of stories as we go, particularly [Post #7](#) (for a simple example of wanting to eat cake) and [Post #13](#) (for the trickier case of social instincts).

3.2.1 Each subsystem generally needs its own sensory processor

For example, in the case of vision, the Steering Subsystem has its superior colliculus, while the Learning Subsystem has its visual cortex. For taste, the Steering Subsystem has its gustatory nucleus of the medulla, while the Learning Subsystem has its gustatory cortex. Etc.

Isn’t that redundant? Some people think so! The book [Accidental Mind](#) by David Linden cites the existence of two sensory-processing systems as a beautiful example of kludgy brain design resulting from evolution’s lack of foresight. But I disagree. They’re not redundant. If I were making an AGI, I would *absolutely* put in two sensory-processing systems!

Why? Suppose that Evolution wants to build a reaction circuit where a genetically-hardwired sensory cue triggers a genetically-hardwired response. For example, as mentioned above, if you’re a mouse, then an expanding dark blob in the upper field-of-view often indicates an incoming bird, and therefore the mouse genome hardwires an expanding-dark-blob-detector to a running-away behavioral circuit.

And I claim that, when building this reaction, the genome *cannot use the visual cortex as its expanding-dark-blob-detector*. Why not? Remember [the previous post](#): the visual cortex learns

from scratch! It takes unstructured visual data and builds a predictive model around it. You can (loosely) think of the visual cortex as a scrupulous cataloguer of patterns in the inputs, and of patterns in the patterns in the inputs, etc. One of these patterns might correspond to expanding dark blobs in the upper field-of-view. Or maybe not! And even if one does, the genome doesn't know in advance *which precise neurons* will be storing that particular pattern. And thus, the genome cannot hardwire those neurons to the running-away behavioral controller.

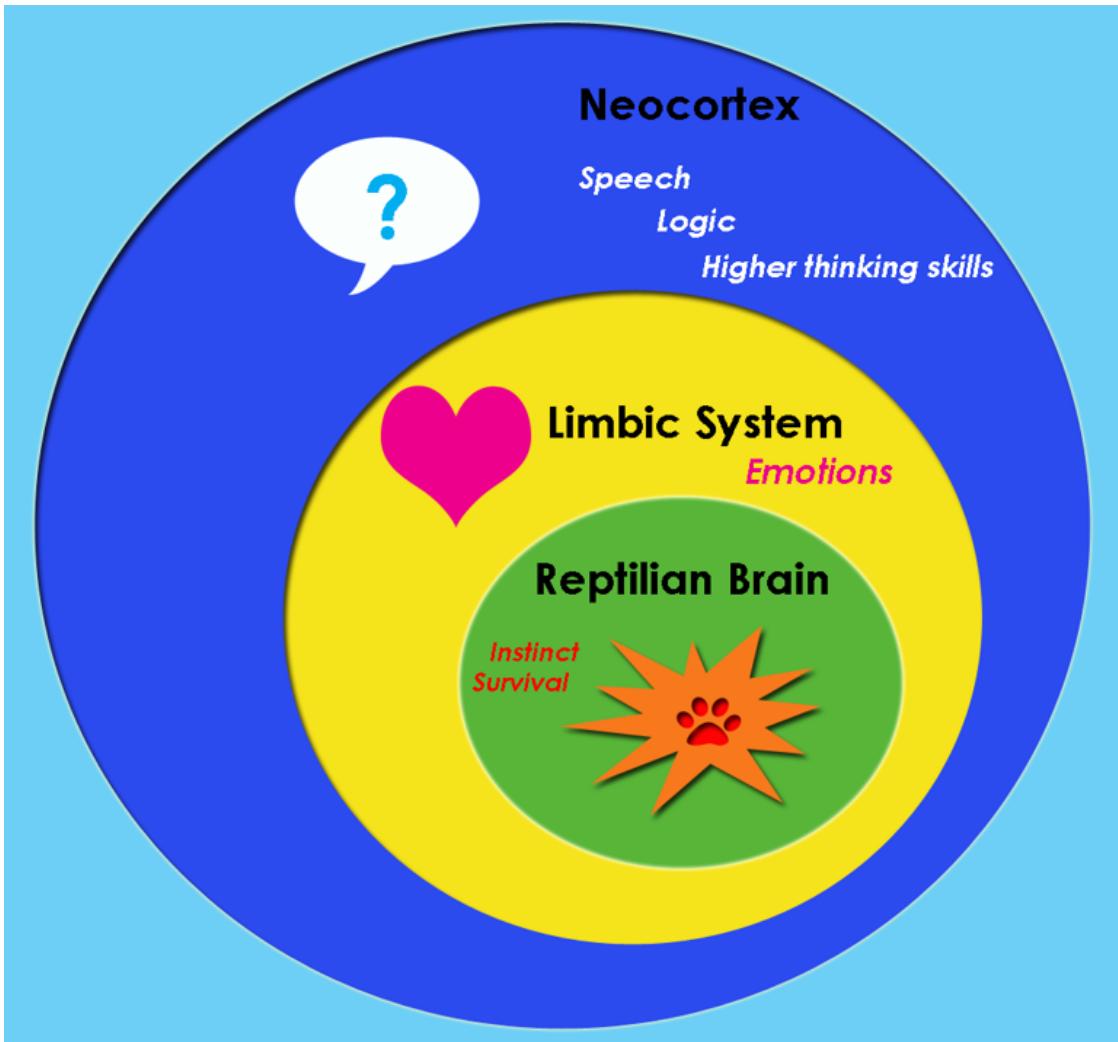
So in summary:

- *Building sensory processing into the Steering Subsystem is a good idea*, because there are lots of areas where it's highly adaptive to attach a genetically-hardwired sensory cue to a corresponding reaction. In the human case, think of fear-of-heights, fear-of-snakes, aesthetics-of-potential-habitats, aesthetics-of-potential-mates, taste-of-nutritious-food, sound-of-screaming, feel-of-pain, and on and on.
- *Building sensory processing into the Learning Subsystem is also a good idea*, because using learning-from-scratch algorithms to learn arbitrary predictive patterns in sensory input within a lifetime is, well, a *really good idea*. After all, many useful sensory patterns are hyper-specific—e.g. “the smell of this one specific individual tree”—such that a corresponding hardwired sensory pattern detector could not have evolved.

Thus, the brain's two sensory-processing systems is *not* an example of kludgy design. It's an example of Orgel's Second Rule: “evolution is cleverer than you are”!

3.3 “Triune Brain Theory” is wrong, but let’s not throw out the baby with the bathwater

In the 1960s & 70s, Paul MacLean & Carl Sagan invented and popularized an idea called the [Triune Brain](#). According to this theory, the brain consists of three layers, stacked on top of each other like an ice cream cone, and which evolved in sequence: first the “lizard brain” (a.k.a. “old brain” or “reptilian brain”) closest to the spinal cord (consisting of the brainstem and basal ganglia); second the “limbic system” wrapped around that (consisting of the amygdala, hippocampus, and hypothalamus), and finally, layered on the outside, the neocortex (a.k.a. “new brain”)—the pièce de résistance, the pinnacle of evolution, the home of human intelligence!!!



The (bad!) triune brain model ([image source](#)).

Well, it's by now well known that **Triune Brain Theory is rubbish**. It lumps brain parts in a way that makes neither functional nor embryological sense, and the evolutionary story is profoundly wrong. For example, half a billion years ago, the earliest vertebrates already had the precursors of *all three* layers of the triune brain—including a “pallium” which would eventually (in our lineage) segregate into the neocortex, hippocampus, part of the amygdala, etc. ([ref](#)).

So yeah, Triune Brain Theory is rubbish. But I freely admit: the story I like (previous section) kinda *rings* of triune brain theory. My Steering Subsystem looks suspiciously like MacLean’s “reptilian brain”. My Learning Subsystem looks suspiciously like MacLean’s “limbic system and neocortex”. MacLean & I have some disagreements about exactly what goes where, and whether the ice cream cone has two scoops versus three. But there’s definitely a resemblance.

My two-subsystem story in this post is not original. You’ll hear a similar story from [Jeff Hawkins](#), [Dileep George](#), [Elon Musk](#), and others.

But those other people tell this story *in the tradition of triune brain theory*, and in particular keeping its problematic aspects, like the “old brain” and “new brain” terminology.

There’s no need to do that!! We can keep the two-subsystem story, while throwing out the triune brain baggage.

So my story is: I think that half a billion years ago, the earliest vertebrates had a (simpler!) learning-from-scratch algorithm in their (proto) telencephalon, and it was “steered” by supervisory signals from their (simpler, proto) brainstem and hypothalamus.

Indeed, we can go back even earlier than vertebrates! There seems to be a homology between the learning-from-scratch cortex in humans and the learning-from-scratch “mushroom body” in fruit flies! ([Further discussion here](#).) I note, for example, that in fruit flies, [odor signals go to both the mushroom body and the lateral horn](#), in beautiful agreement with the general principle that sensory inputs need to go to both the Learning Subsystem and the Steering Subsystem (Section 3.2.1 above).

Anyway, in the 700 million years since our last common ancestor with insects, *both* the Learning Subsystem *and* the Steering Subsystem have dramatically expanded and elaborated in our lineage.

But that doesn’t mean that they contribute equally to “human intelligence”. Again, both are essential, but I think it’s strongly suggestive that ~96% of human brain volume is the Learning Subsystem. Focusing more specifically on the telencephalon part (which includes the neocortex in mammals), its fraction of brain volume is 87% in humans ([ref](#)), 79% in chimps ([ref](#)), 77% in certain parrots, 51% in chickens, 45% in crocodiles, and just 22% in frogs ([ref](#)). There’s an obvious pattern here, and I think it’s right: namely, that to get recognizably intelligent and flexible behavior, you need a massively-scaled-up Learning Subsystem.

See? I can tell my two-subsystem story with none of that “old brain, new brain” nonsense.

3.4 Three types of ingredients in a Steering Subsystem

I’ll start with the summary table, and then elaborate on it in the following subsections.

3.4.1 Summary table

| Category of Steering Subsystem ingredient | Possible examples | Present in (competent) humans? | Expected in future AGIs? |
|---|--|--------------------------------|--------------------------|
| (A) Things the Steering Subsystem <i>needs</i> to do in order to get general intelligence | <ul style="list-style-type: none">• Curiosity drive (?)• Drive to attend to certain types of things in the environment (humans, language, technology, etc.) (?)• General involvement in helping establish the Learning Subsystem neural architecture (?) | Yes, by definition | Yes |
| (B) Everything else in | | Often, but not | |

| | | | |
|---|---|--|--|
| a neurotypical human's Steering Subsystem | <ul style="list-style-type: none"> • Social instincts (which underlie altruism, love, remorse, guilt, sense-of-justice, loyalty, etc.) • Drives underlying disgust, aesthetics, transcendence, serenity, awe, hunger, pain, fear-of-spiders, etc. | always—for example, high-functioning sociopaths seem to be missing some of the usual social instincts. | Not “by default”, but it’s <i>possible</i> if we: <ol style="list-style-type: none"> (1) figure out exactly how they work, and (2) convince AGI developers to put them in. |
| (C) Every other possibility, most of which are <i>completely unlike anything</i> in the Steering Subsystem of humans or indeed any animal | <ul style="list-style-type: none"> • Drive to increase a company’s bank account balance? • Drive to invent a better solar cell? • Drive to do whatever my human supervisor wants me to do? (<i>There’s a catch: no one knows how to implement this one!</i>) | No | Yes “by default”. If something is a bad idea, we can try to convince AGI developers not to do that. |

3.4.2 Aside: what do I mean by “drives”?

I’ll elaborate on this picture in later posts, but for now let’s just say that the Learning Subsystem does reinforcement learning (among other things), and the Steering Subsystem sends it rewards. The components of the reward function relate to what I’ll call “innate drives”—they’re the root cause of why some things are inherently motivating / appetitive and other things are inherently demotivating / aversive.

Explicit goals like “I want to get out of debt” are different from innate drives. Explicit goals come out of a complicated dance between “innate drives in the Steering Subsystem” and “learned content in the Learning Subsystem”. Again, much more on that topic in future posts.

Remember, innate drives are in the Steering Subsystem, whereas the abstract concepts that make up your conscious world are in the Learning Subsystem. For example, if I say something like “altruism-related innate drives”, you need to understand that I’m *not* talking about “the abstract concept of altruism, as defined in an English-language dictionary”, but rather “some innate Steering Subsystem circuitry which is *upstream* of the fact that neurotypical people sometimes find altruistic actions to be inherently motivating”. There is *some* relationship between the abstract concepts and the innate circuitry, but it might be a complicated one—nobody expects a one-to-one relation between N discrete innate circuits and a corresponding set of N English-language words describing emotions and drives.^[1]

With that out of the way, let’s move on to more details about that table above.

3.4.3 Category A: Things the Steering Subsystem needs to do in order to get general intelligence (e.g. curiosity drive)

Let's start with the "**curiosity drive**". If you're not familiar with the background of "curiosity" in ML, I recommend [The Alignment Problem by Brian Christian](#), chapter 6, which contains the gripping story of how researchers eventually got RL agents to win the Atari game *Montezuma's Revenge*. Curiosity drives seem essential to good performance in ML, and humans also seem to have an innate curiosity drive. I assume that future AGI algorithms will need a curiosity drive as well, or else they just won't work.

To be more specific, I think this is a bootstrapping issue—I think we need a curiosity drive early in training, but can probably turn it off eventually. Specifically, let's say there's an AGI that's generally knowledgeable about the world and itself, and capable of getting things done, and right now it's trying to invent a better solar cell. I claim it probably doesn't need to feel an innate curiosity drive. Instead it may seek new information, and seek surprises, *as if* it were innately curious, because it has learned through experience that seeking those things tends to be an effective strategy for inventing a better solar cell. In other words, something like curiosity can be *motivating as a means to an end*, even if it's not *motivating as an end in itself*—curiosity can be a learned metacognitive heuristic. See [instrumental convergence](#). But that argument does not apply early in training, when the AGI starts from scratch, knowing nothing about the world or itself. Instead, early in training, I think we really need the Steering Subsystem to be holding the Learning Subsystem's hand, and pointing it in the right directions, if we want AGI.

Another possible item in Category A is an **innate drive to pay attention to certain things in the environment, e.g. human activities, or human language, or technology**. I don't know *for sure* that this is necessary, but it seems to me that a curiosity drive *by itself* wouldn't do what we want it to do. It would be completely undirected. Maybe it would spend eternity running [Rule 110](#) in its head, finding deeper and deeper patterns, while completely ignoring the physical universe. Or maybe it would find deeper and deeper patterns in the shapes of clouds, while completely ignoring everything about humans and technology. In the human brain case, the human brainstem definitely has a mechanism for forcing attention onto human faces ([ref](#)), and I strongly suspect that there's a system that forces attention onto human speech sounds as well. I could be wrong, but my hunch is that something like that will need to be in AGIs too. As above, if this drive is necessary at all, it might only be necessary early in training.

What else might be in Category A? On the table above, I wrote the vague "General involvement in helping establish the Learning Subsystem neural architecture". This includes sending reward signals and error signals and hyperparameters etc. to particular parts of the neural architecture in the Learning Subsystem. For example, in [Post #6](#) I'll talk about how only *part* of the neural architecture gets the main RL reward signal. I think of these things as (one aspect of) how the Learning Subsystem's neural architecture is actually implemented. AGIs will have some kind of neural architecture too, although maybe not exactly the same as humans'. Therefore, they might need some of these same kinds of signals. I talked about neural architecture briefly in [Section 2.8 of the last post](#), but mostly it's irrelevant to this series, and I won't talk about it beyond this unhelpfully-vague paragraph.

There might be other things in Category A that I'm not thinking of.

3.4.4 Category B: Everything else in the human Steering Subsystem (e.g. altruism-related drives)

I'll jump right into what I think is most important: **social instincts**, including various drives related to altruism, sympathy, love, guilt, remorse, status, jealousy, sense-of-fairness, etc. Key question: **How do I know that social instincts belong here in Category B, i.e. that they aren't one of the Category A things that are essential for general intelligence?**

Well, for one thing, look at high-functioning sociopaths. I've had the unfortunate experience of getting to know a couple of them very well in my day. They understood the world, and themselves, and language and math and science and technology, and they could make elaborate plans to successfully accomplish impressive feats. If there were an AI that could do everything that a high-functioning sociopath can do, we would *unhesitatingly* call it "AGI". Now, I think high-functioning sociopaths have *some* social instincts—they're more interested in manipulating people than manipulating toys—but their social instincts seem to be *very different* from those of a neurotypical person.

Then on top of that, we can consider people with autism, and people with schizophrenia, and [SM](#) (who is missing her amygdala and more-or-less lacks negative social emotions), and on and on. All these groups of people have "general intelligence", but their social instincts / drives are all quite different from each other's.^[2]

All things considered, I find it very hard to believe that any aspect of social instincts is essential for general intelligence. I think it's at least open to question whether social instincts are even *helpful* for general intelligence!! For example, if you look at the world's most brilliant scientific minds, I'd guess that people with neurotypical social instincts are if anything slightly *underrepresented*.

One reason this matters is that, I claim, **social instincts underlie "the desire to behave ethically"**. Again, consider high-functioning sociopaths. They can *understand* honor and justice and ethics if they try—in the sense of correctly answering quiz questions about what is or isn't honorable etc.—they're just not *motivated* by it.^[3]

If you think about it, it makes sense. Suppose I tell you "You really ought to put pebbles in your ears." You say "Why?" And I say "Because, y'know, your ears, they don't have any pebbles in them, but they really should." And again you say "Why?" ...At some point, this conversation has to ground out at something that you find *inherently* motivating or demotivating, in and of itself. And I claim that social instincts—the various innate drives related to sense-of-fairness and sympathy and loyalty and so on—are ultimately providing the ground on which those intuitions stand.

(I'm not taking a stand on moral realism vs. moral relativism here—i.e., the question of whether there is a "fact of the matter" about what is ethical vs. unethical. Instead, I'm saying that *if* there's an agent that is completely lacking in any innate drives that might spur a desire to act ethically, then then we can't expect the agent to act ethically, no matter how intelligent and capable it is. Why would it? Granted, it might act ethically *as a means to an end*—e.g. to win allies—but that doesn't count. More discussion and intuition-pumps in [my comment here](#).)

That's all I want to say about social instincts for now; I'll return to them in [Post #13](#).

What else goes in Category B? Lots of things!! There's disgust, and aesthetics, and transcendence, and serenity, and awe, and hunger, and pain, and fear-of-spiders, etc.

3.4.5 Category C: Every other possibility (e.g. drive to increase my bank account balance)

When people make AGIs, they can put *whatever they want* into the reward function! This would be analogous to inventing new innate drives out of whole cloth. And these can be innate drives that are radically unlike anything in humans or animals.

Why might the future AGI programmers invent new-to-the-world innate drives? Because it's the obvious thing to do!! Go kidnap a random ML researcher from the halls of NeurIPS, drive them to an abandoned warehouse, and force them to make a bank-account-balance-increasing AI using reinforcement learning.^[4] I bet you anything that, when you look at their source code, you're going to find a reward function that involves the bank account balance. You won't find anything like *that* among the genetically-hardwired circuitry in the human brainstem! It's a new-to-the-world innate drive.

Not only is “put in an innate drive for increasing the bank account balance” the obvious thing to do, but I think it would actually work! For a while! And then it would fail catastrophically! It would fail as soon as the AI became competent enough to find out-of-the-box strategies to increase the bank account balance—like borrowing money, hacking into the bank website, and so on.

(Related: [hilarious and terrifying list of historical examples of AIs finding unintended, out-of-the-box strategies for maximizing a reward](#). More on this in future posts.) In fact, this bank-account-balance example is one of the many, many possible drives that would plausibly lead to an AGI harboring a secret motivation to escape human control and kill everyone (see [Post #1](#)).

So these kinds of motivations are the worst: they’re dangling right in front of everyone’s faces, they’re the best way to get things done and publish papers and beat benchmarks if the AGI is not overly clever, and then when the AGI becomes competent enough, they lead to catastrophic accidents.

Maybe you’re thinking: “It’s *really obvious* that an AGI with an all-consuming innate drive to increase a certain bank account balance is an AGI that would try to escape human control, self-reproduce etc. Do you really believe that future AGI programmers would be *so reckless* as to put in something like that??”

Well, umm, yes. Yes, I do. But even setting that aside for the sake of argument, there’s a bigger problem: we don’t currently know how to code up *any innate drive whatsoever* such that the resulting AGI would definitely stay under control. Even the drives that *sound* benign are probably not, at least not in our current state of knowledge. Much more on this in later posts (especially [#10](#)).

To be sure, Category C is a very big tent. I would not be at all surprised if there exist Category C innate drives that would be *very good* for AGI safety! We just need to find them! I’ll be exploring this design space later in the series.

3.5 Brain-like AGI will by default have radically nonhuman (and dangerous) motivations

I mentioned this way back in the [first post \(Section 1.3.3\)](#), but now we have the explanation.

The previous subsection proposes three types of ingredients to put in a Steering Subsystem: (A) Those necessary to wind up with an AGI at all, (B) Everything else in humans, (C) Anything *not* in humans.

My claims are:

1. People want to make powerful AIs with state-of-the-art capabilities in challenging domains—they know that it’s good for publications, good for impressing their colleagues, getting jobs and promotions and grants, etc. I mean, just look at AI and ML today. Therefore, by default, I expect AGI researchers to race down the most direct path to AGI: reverse-engineering the Learning Subsystem, and combining it with Category-A drives.
2. Category B contains some drives that are plausibly useful for AGI safety: drives related to altruism, sympathy, generosity, humility, etc. Unfortunately, we don’t currently know how any of those drives are implemented in the brain. And figuring that out is unnecessary for building AGIs. So by default, I think we should expect AGI researchers to ignore Category B until they have AGIs up and running, and *only then* start scrambling to figure out how to build altruism drives etc. And they might outright fail—it’s totally possible that the corresponding brainstem & hypothalamus circuitry is a frightfully complicated mess, and we only have so much time between “AGIs are up and running” and “someone accidentally makes an out-of-control AGI that kills everyone” (see [Post #1](#)).
3. There are things in Category C like “*A low-level innate drive to increase a particular bank account balance*” that are immediately obvious to everyone, and easy to implement, and

will work well at accomplishing the programmers' goals *while their janky proto-AGIs are not yet very capable*. Therefore, by default, I expect future researchers to use these kinds of "obvious" (but dangerous and radically-nonhuman) drives as they work towards developing AGI. And as discussed above (and more in later posts), even if the researchers start trying in good faith to give their AGI an innate drive for being helpful / docile / whatever, they might find that they don't know how to do so.

In sum, if researchers travel down the most easy and natural path—the path that looks like the AI and neuroscience R&D community continuing to behave in ways that they behave right now—we will wind up being able to make AGIs that do impressive things that their programmers want, for a while, but are driven by radically alien motivation systems that are fundamentally unconcerned with human welfare, and these AGIs will try to escape human control as soon as they are capable enough to do so.

Let's try to change that! In particular, if we can figure out *in advance* how to write code that builds an innate drive for altruism / helpfulness / docility / whatever, that would be a huge help. This will be a major theme of this series. But don't expect final answers. It's an unsolved problem; there's still a lot of work to do.

3.6 Response to Jeff Hawkins's argument against AGI accident risk

Jeff Hawkins has a recent book *A Thousand Brains*. I wrote a more detailed book review [here](#). Jeff Hawkins is a strong advocate of a two-subsystems perspective very similar to mine. No coincidence—his writings helped push me in that direction!

To Hawkins's great credit, he takes ownership of the idea that his neuroscience / AI work is pushing down a path (of unknown length) towards AGI, and he has tried to think carefully about the consequences of that larger project—as opposed to the more typical perspective of declaring AGI to be someone else's problem.

So, I'm delighted that Hawkins devotes a large section of his book to an argument about AGI catastrophic risk. But his argument is *against* AGI catastrophic risk!! What's the deal? How do he and I, starting from a similar two-subsystems perspective, wind up with diametrically opposite conclusions?

Hawkins makes many arguments, and again I addressed them more comprehensively in [my book review](#). But here I want to emphasize two of the biggest issues that bear on this post.

Here's my paraphrase of a particular Hawkins argument. (I'm translating it into the terminology I'm using in this series, e.g. he says "old brain" where I say "Steering Subsystem". And maybe I'm being a bit mean. You can read the book and judge for yourself whether this is fair.)

1. The Learning Subsystem (neocortex etc.) *by itself* has no goals or motivations. It won't do anything. It certainly won't do anything dangerous. It's like a map sitting on a table.
2. Insofar as humans have problematic drives (greed, self-preservation, etc.), they come from the Steering Subsystem (brainstem etc.).
3. The thing that I, Jeff Hawkins, am proposing, and doing, is trying to reverse-engineer the Learning Subsystem, not the Steering Subsystem. So what the heck is everyone so worried about?
4. ...
5. ...
6. Oh hey, on a *completely* unrelated note, we will eventually make future AGIs, and these will have not only a Learning Subsystem, but also a Steering Subsystem attached to it. I'm not going to talk about how we'll design the Steering Subsystem. It's not really something that I think about much.

Each of these points *in isolation* seems reasonable enough. But when you put them together, there's a gaping hole! Who cares if a neocortex *by itself* is safe? A neocortex *by itself* was never the plan! The question we need to ask is whether an AGI consisting of *both* subsystems attached together will be safe. And that depends crucially on how we build the Steering Subsystem. Hawkins isn't interested in that topic. But I am! Read on in the series for much more on this. [Post #10](#) in particular will dive into why it's a heck of a lot harder than it sounds to build a Steering Subsystem that steers the AGI into doing some particular thing that we intend for it to do, without also incidentally instilling dangerous antisocial motivations that we never intended it to have.

One more (related) issue that I didn't mention in my earlier book review: I think that Hawkins is partly driven by an intuition that I argued against in ([Brainstem, Neocortex](#)) ≠ ([Base Motivations, Honorable Motivations](#)). (and more on that topic coming up in [Post #6](#)): a tendency to inappropriately locate ego-syntonic motivations like "unraveling the secrets of the universe" in the neocortex (Learning Subsystem), and ego-dystonic motivations like hunger and sex drive in the brainstem (Steering Subsystem). I claim that the correct answer is that *all* motivations come ultimately from the Steering Subsystem, no exceptions. This will hopefully be obvious if you keep reading this series.

In fact, my claim is even implied by the better parts of Hawkins's own book! For example:

- Hawkins in Chapter 10: "The neocortex learns a model of the world, which by itself has no goals or values."
- Hawkins in Chapter 16: " 'We'—the intelligent model of ourselves residing in the neocortex—are trapped. We are trapped in a body that...is largely under the control of an ignorant brute, the old brain. We can use intelligence to imagine a better future.... But the old brain could ruin everything..."

To spell out the contradiction: if "we" = the neocortex's model, and the neocortex's model has no goals or values whatsoever, then "we" certainly would not be aspiring to a better future and hatching plots to undermine the brainstem.

3.7 Timelines-to-brain-like-AGI part 2 of 3: how hard will it be to reverse-engineer the Steering Subsystem well enough for AGI?

(Reminder: Timelines Part 1 of 3 was [Section 2.8 of the previous post](#).)

Above (Section 3.4.3), I discussed “Category A”, the minimal set of ingredients to build an AGI-capable Steering Subsystem (not necessarily *safe*, just *capable*).

I don’t *really* know what is in this set. I suggested that we’d probably need some kind of curiosity drive, and maybe some drive to pay attention to human language and other human activities, and maybe some signals that go along with and help establish the Learning Subsystem’s neural network architecture.

If that’s right, well, this doesn’t strike me as too hard! Certainly it’s a *heck* of a lot easier than reverse-engineering everything in the human hypothalamus and brainstem! Keep in mind that there is a substantial literature on curiosity in both ML ([1](#), [2](#)) and psychology. “A drive to pay attention to human language” requires nothing more than a classifier that says (with reasonable accuracy, it doesn’t have to be perfect) whether any given audio input is or isn’t human language; that’s *trivial* with today’s tools, if it’s not already on GitHub.

I think we should be open to the possibility that it just isn’t that hard to build a Steering Subsystem that (together with a reverse-engineered Learning Subsystem, see [Section 2.8 of the previous post](#)) can develop into an AGI after training. Maybe it’s not decades of R&D; maybe it’s not even years of R&D! Maybe a competent researcher will nail it after just a couple tries. On the other hand—maybe not! Maybe it *is* super hard! I think it’s very difficult to predict how long it would take, from our current vantage point, and that we should remain uncertain.

3.8 Timelines-to-brain-like-AGI part 3 of 3: scaling, debugging, training, etc.

Having a fully-specified, AGI-capable algorithm isn't the end of the story; you still need to implement the algorithm, iterate on it, hardware-accelerate and parallelize it, work out the kinks, run trainings, etc. We shouldn't *ignore* that part, but we shouldn't overstate it either. I won't get into this here, because I recently wrote a whole separate blog post about it:

[Brain-inspired AGI and the “lifetime anchor”](#)

The upshot of that post is: I think all that stuff could absolutely get done in <10 years. Maybe even <5. Or it could take longer. I think we should be very uncertain.

Thus concludes my timeline-to-brain-like-AGI discussion, which again is not my main focus in this series. You can read my three timelines sections ([2.8](#), 3.7, and this one), agree or disagree, and come to your own conclusions.

3.9 Timelines-to-brain-like-AGI encore: How should I *feel* about a probabilistic timeline?

My “timelines” discussion (Sections [2.8](#), 3.7, 3.8) has been about the *forecasting* question “what probability distribution should I assign to when AGI will arrive (if ever)?”

Semi-independent of that question is a kind of *attitude* question: “How should I *feel* about that probability distribution?”

For example, there can be two people who *both* agree with (just an example) “35% chance of AGI by 2042”. But their *attitudes* may be wildly different:

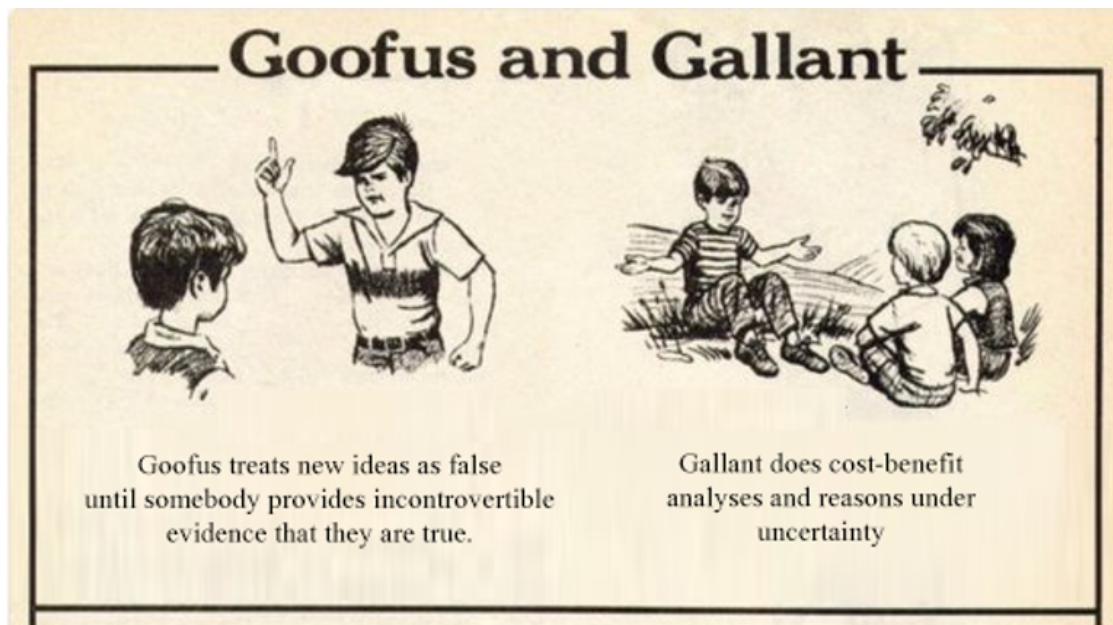
- One of the two people rolls their eyes, laughs, and says: “See, I told you! AGI probably isn’t coming for *decades*!”
- The other person widens their eyes, drops their jaw, and says “Oh. My. God. Excuse me for a moment while I rethink everything about my life.”

There are a lot of factors underlying these different attitudes towards the same belief about the world. First, some factors are kinda more questions of psychology rather than questions of fact:

- “What attitude better fits my self-image and psychology?”—ooh, yikes, this one cuts deep into our psyches. People who think of themselves as cool-headed serious skeptical dignified grounded scientists may feel irresistibly drawn to the belief that AGI isn’t a big deal. People who think of themselves as pioneering radical transhumanist technologists may equally feel irresistibly drawn to the opposite belief that AGI will radically change *everything*. I bring this up so that you can meditate on your own biases. Oh, who am I kidding; realistically, I just handed you a nice way to smugly mock and dismiss anyone who disagrees with you. (You’re welcome!) For my part, I claim some immunity to being dismissed-via-psychoanalysis: When I first came to believe that AGI is a very big deal, I *totally* self-identified as a cool-headed serious skeptical dignified grounded middle-aged scientist, with no interest in, nor connection to, science fiction or transhumanism or the tech industry or AI or silicon valley, etc. Take that! Ha! But really, this is a stupid game to play: dismissing people’s beliefs by psychoanalyzing them for hidden motives has always been a terrible idea. It’s too easy. Right or wrong, you can always find a good reason to smugly question the motives of anyone you disagree with. It’s just a cheap trick to avoid the hard work of figuring out whether they might actually be correct. Also on the general

topic of psychology: taking our possible AGI future seriously (as seriously as I think is warranted) can be, well, kinda wrenching! It was hard enough getting used to the idea that Climate Change is really happening, right?? See [this post](#) for more on that.

- How should I think about possible-but-uncertain future events? I suggest reading [this Scott Alexander post](#). Or if you prefer the meme version:



[Image source: Scott Alexander.](#)

Relatedly, there's a kind of feeling expressed by [the famous "Seeing the Smoke" essay](#), and this meme here:



Loosely based on a [@Linch](#) meme, if I recall correctly.

To spell it out, the *right* idea is to weigh risks and benefits and probabilities of over-preparing vs. under-preparing for an uncertain future risk. The *wrong* idea is to add an extra entry into that ledger—"the risk of looking foolish in front of my friends by over-preparing for something weird that winds up not being a big deal"—and treat that one entry as *overwhelmingly more important than everything else on the list*, and then it follows that we shouldn't try to mitigate a possible future catastrophe until we're >99.9% confident that the catastrophe will definitely happen, in a

kind of insane bizarro-world reversal of Pascal's Wager. Luckily, this is increasingly a moot point; your friends are less and less likely to think you're weird, because AGI safety has gotten much more mainstream in recent years—thanks especially to outreach and pedagogy by [Stuart Russell](#), [Brian Christian](#), [Rob Miles](#), and many others. You can help that process along by sharing this post series! ;)

Putting those aside, other reasons for different attitudes towards AGI timelines are more substantive, particularly the questions:

- How much will AGI transform the world? For my part, I'm way the heck over on the "lots" end of the spectrum. I endorse the Eliezer Yudkowsky [quote](#): "Asking about the effect of [superhuman AGI] on [unemployment] is like asking how US-Chinese trade patterns would be affected by the Moon crashing into the Earth. There would indeed be effects, but you'd be missing the point." For a more sober discussion, try Holden Karnofsky's [Digital People Would Be An Even Bigger Deal](#), and maybe also [This Can't Go On](#) as background, and what the heck, [the whole rest of that series too](#). Also see [here](#) for some numbers suggesting that brain-like AGI will probably *not* require so many computer chips or so much electricity that it can't be widely used.
- How much do we need to do, to prepare for AGI? See [Post #1, Section 1.7](#) for my argument that we're way behind schedule, and later in this series I'll be discussing the many still-unsolved problems.

1. ^

Well, maybe *some* people expect that there's a one-to-one correspondence between English-language abstract concepts like "sadness" and corresponding innate reactions. If you read the book [How Emotions Are Made](#), Lisa Feldman Barrett spends hundreds of pages belaboring this point. She must have been responding to *somebody*, right? I mean, it feels to me like an absurd straw-man to say "Each and every situation that a native English speaker would describe as 'sadness' corresponds to the exact same innate reaction with the exact same facial expression." I'd be surprised if even [Paul Ekman](#) (whom Barrett was supposedly rebutting) actually believes that, but I dunno.

2. ^

I wouldn't suggest that the Steering Subsystem circuitry underlying social instincts is built in a fundamentally different way in these different groups—that would be evolutionarily implausible. Rather, I think there are lots of adjustable parameters on how strong the different drives are, and they can be set to wildly different values, including the possibility that a drive is set to be so weak as to be effectively absent. See my speculation on autism and psychopathy [here](#).

3. ^

See Jon Ronson's *The Psychopath Test* for a fun discussion of attempts to teach empathy to psychopaths. The students merely wound up better able to *fake* empathy in order to manipulate people. Quote from one person who taught such a class: "I guess we had inadvertently created a finishing school for them."

4. ^

I suppose I could have *hired* an ML researcher instead. But who could afford the salary?

How I Formed My Own Views About AI Safety

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

This is a linkpost for <https://www.neelnanda.io/blog/47-inside-views>

Disclaimer: I work as a researcher at Anthropic, but this post entirely represents my own views, rather than the views of my own employer

Introduction

I've spent the past two years getting into the field of AI Safety. One important message I heard as I was entering the field was that I needed to "form an inside view about AI Safety", that I needed to form my own beliefs and think for myself rather than just working on stuff because people smarter than me cared about it. And this was incredibly stressful! I think the way I interpreted this was pretty unhealthy, caused me a lot of paralysing uncertainty and anxiety, and almost caused me to give up on getting into the field. But I feel like I've now reached a point I'm comfortable with, and where I somewhat think I have my own inside views on things and understand how to form them.

In this post, I try to explain the traps I fell into and why, what my journey actually looked like, and my advice for how to think about inside views, now I've seen what *not* to do. This is a complex topic and I think there are a lot of valid perspectives, but hopefully my lens is novel and useful for some people trying to form their own views on confusing topics (AI Safety or otherwise)! (Note: I don't discuss why I *do* now think AI Safety is important and worth working on - that's a topic for a future post!)

The Message of Inside Views

First, context to be clear about what I mean by **inside views**. As I understand it, this is a pretty fuzzily defined concept, but roughly means "having a clear model and argument in my head, starting from some basic and reasonable beliefs about the world, that get to me to a conclusion like 'working on AI Safety is important' without needing to rely on deferring to people". This feels highly related to the concept of [gears-level models](#). This is in comparison to **outside views**, or **deferring** to people, where the main reason I believe something is because smart people I respect believe it. In my opinion, there's a general vibe in the rationality community that inside views are good and outside views are bad (see Greg Lewis' [In Defence of Epistemic Modesty](#) for a good argument for the importance of outside views and deferring!). Note that this is *not* the Tetlockian sense of the words, used in forecasting, where outside view means 'look up a base rate' and inside view means 'use my human intuition, which is terribly calibrated', where the standard wisdom is outside view > inside view.

Good examples of this kind of reasoning: Buck Shlegeris' [My Personal Cruxes for Working on AI Safety](#), Richard Ngo's [AGI Safety from First Principles](#), Joseph Carlsmith's report on [Existential Risk from Power-Seeking AI](#). Note that, while these are all about the question of 'is AI Safety a problem at all', the notion of an inside view

also applies well to questions like ‘de-confusion research/reinforcement learning from human feedback/interpretability is the best way to reduce existential risk from AI’, arguing for specific research agendas and directions.

How I Interpreted the Message of Inside Views

I’m generally a pretty anxious person and bad at dealing with uncertainty, and sadly, this message resulted in a pretty unhealthy dynamic in my head. It felt like I had to figure out for myself the conclusive truth of ‘is AI Safety a real problem worth working on’ and which research directions were and were not useful, so I could then work on the optimal one. And that it was my responsibility to do this all myself, that it was bad and low-status to work on something because smart people endorsed it.

This was hard and overwhelming because there are a *lot* of agendas, and a lot of smart people with different and somewhat contradictory views. So this felt basically impossible. But it *also* felt like I had to solve this before I actually started any permanent research positions (ie by the time I graduated) in case I screwed up and worked on something sub-optimal. And thus, I had to solve this problem that empirically most smart people must be screwing up, and do it all before I graduated. This seemed basically impossible, and created a big [ugh field](#) around exploring AI Safety. Which was already pretty aversive, because it involved re-skilling, deciding between a range of different paths like PhDs vs going straight into industry, and generally didn’t have a clean path into it.

My Journey

So, what actually happened to me? I started taking AI Safety seriously in my final year of undergrad. At the time, I bought the heuristic arguments for AI Safety (like, something smarter than us is scary), but didn’t really know what working in the field looked like beyond ‘people at MIRI prove theorems I guess, and I know there are people at top AI labs doing safety stuff?’ I started talking to lots of people who worked in the field, and gradually got data on what was going on. This was all pretty confusing and stressful, and was competing with going into quant finance - a safe, easy, default path that I already knew I’d enjoy.

After graduating, I realised I had a lot more flexibility than I thought. I took a year out, and managed to finagle my way into doing three back-to-back AI Safety internships. The big update was that I could explore AI Safety without risking too much - I could always go back into finance in a year or two if it didn’t work out. I interned at FHI, DeepMind and CHAI - working on mathematical/theoretical safety work, empirical ML based stuff to do with fairness and bias, and working on empirical interpretability work respectively. I also did the AGI Fundamentals course, and chatted to a lot of people at the various orgs I worked at and at conference. I tried to ask all the researchers I met about their theory of change for how their research actually matters. One thing that really helped me was chatting to a researcher at OpenAI who said that, when he started, he didn’t have clear inside views. But that he’d formed them fairly organically over time, and just spending time thinking and being in a professional research environment was enough.

At the end of the year, I had several offers and ended up joining Anthropic to work on interpretability with Chris Olah. I wasn’t sure this was the best option, but I was really

excited about interpretability, and it seemed like the best bet. A few months in, this was clearly a *great* decision and I'm really excited about the work, but it wouldn't have been the end of the world if I'd decided the work wasn't very useful or a bad fit, and I expect I could have left within a few months without hard feelings. As I've done research and talked to Chris + other people here, I've started to form clearer views on what's going on with interpretability and the theory of impact for it and Anthropic's work, but there's still big holes in my understanding where I'm confused or deferring to people. And this is fine! I don't think it's majorly holding me back from having an impact in the short-term, and I'm forming clearer views with time.

My Advice for Thinking About & Forming Inside Views

Why to form them?

I think there are four main reasons to care about forming inside views:

- **Truth-tracking** - having an impact is hard! It's really important to have true beliefs, and the best way to find them is by trying hard to form your own views and ensuring they correlate with truth. It's easy to get deferring wrong if you trust the wrong people.
 - I'm pretty unconvinced by this one - it doesn't seem that hard to find people smarter than me, who've thought about each problem for longer than I have, and just believing whatever they believe. Especially if I average multiple smart people's beliefs
 - Eg, I haven't thought too much about biosecurity, but will happily defer to people like Greg Lewis on the topic!
- **Ensuring good community epistemic health** - Maybe your personal inside view will track the truth less well than the best researchers. But it's not perfectly correlated! If you try hard to find the truth on your own, you might notice ideas other people are missing, can poke holes in popular arguments, etc. And this will make the community as a whole better off
 - This one is pretty legit, but doesn't seem *that* big a deal. Like, important, sure, but not something I'd dedicate more than 5% of my effort towards max
 - It seems particularly important to avoid information cascades where I work on something because Alice thinks it matters, and then Bob is a bit skeptical of Alice alone but observes that both Alice *and* I believe it matters, and works on it even harder, Charlie sees me, Alice and Bob, etc. This is a main reason I try hard to distinguish between what I believe all things considered (including other people's views) and what I believe by my own lights (according to my own intuitions + models of the world)
- **Motivation** - It's really hard to work on something you don't believe in!
 - I personally overthink things, and this one is really important to me! But people vary - this is much more a fact about personal psychology than an abstract statement about how to have an impact
- **Research quality** - Doing good research involves having good intuitions and research taste, sometimes called an inside view, about why the research matters and what's really going on. This conceptual framework guides the many small decisions and trade-offs you make on a daily basis as a researcher

- I think this is really important, but it's worth distinguishing this from 'is this research agenda ultimately useful'. This is still important in eg pure maths research just for doing good research, and there are areas of AI Safety where you can do 'good research' without actually reducing the probability of x-risk.
 - Toy example: Let's say there are ten good AI Safety researchers in the world, who all believe different things. My all-things-considered view should put 10% credence on each person's view. But I'll get *much* more research done if I randomly pick one person and fully adopt their views and dive into their research agenda. So, even if only one researcher is correct, the latter strategy is much better in expected value.
- This is one of the main reasons that mentorship is so key. I have become a way more effective interpretability researcher by having ready access to Chris to ask for advice, intuitions and direction. And one of my top priorities is absorbing as many of his conceptual frameworks as I can
 - More generally, IMO the point of a research mentor is to lend you their conceptual frameworks to advise you on how to make the right decisions and trade-offs. And you slowly absorb their frameworks by supervised learning, and build on and add to them as you grow as a researcher

These are pretty different, and it's really important to be clear about which reasons you care about! Personally, I mostly care about motivation > research quality = impact >> community epistemics

How to form them?

- **Talk to people!** Try to absorb *their* inside views, and make it your own
 - Importantly, the goal is not to defer to them, it's to understand what they believe *and why*.
 - My main tool for this is to ask lots of questions, and then **paraphrase** - summarise back my understanding in my own words, and ask what's wrong or what I'm missing.
 - My default question is 'so, why, concretely, does your research direction reduce existential risk from AI?'
 - Or, 'what are the biggest ways you disagree with other researchers?' Or 'why aren't you working on X?'
 - I really, really love paraphrasing! A few reasons it's great:
 - It forces you to actively listen and process in the moment
 - It's much easier to correct than teach - the other person can easily identify issues in your paraphrase and correct them
 - It makes it obvious to myself if I'm confused or don't understand something, or if I'm deferring on any points - it's awkward to say things that are confused!
 - Once I get it working, I have now downloaded their mental model into my head and can play around with it
 - Once you've downloaded multiple people's models, you can compare them, see how they differ, etc
 - A variant - focus on **cruxes**, key claims where if they changed their mind on that they'd change their mind about what to work on.
 - This is really important - some people work on a direction because they think it's the most important, other people work on it because

eg it's a good personal fit or they find it fun. These should be *completely different* conversations

- A variant - write a google doc summarising a conversation and send it to them afterwards for comments. This can work great if you find it hard to summarise in the moment, and can produce a good artefact to publish or share - I'd love it if people did this more with me
- **You have permission to disagree** (even with really cool and high-status people)
 - This was a big update for me! Someone being smart and competent just means they're right more often, not that they're always right
 - It really helps to have a low bar for asking dumb questions - if you poke at everything that might be wrong, 90% of the time they're right and you learn something, and 10% of the time they missed something
 - For example, I've done research in the past that, in hindsight, I don't think was particularly useful. And this is totally fine!
 - Empirically, there's a lot of smart people who believe different and contradictory things! It's impossible for all of them to be right, so you *must* disagree with some of them. Internalising that you can do this is really important for being able to think clearly
- **Don't be a monk** - you form an inside view by going out in the world and doing things - not just by hiding away and thinking really hard
 - Eg, just try doing research! Spend 10 hours pursuing something, write up a blog post, fail, succeed, hear criticism, see what you learn and make updates
 - Talk to lots of people!
 - Live your life, and see what happens - my thoughts naturally change a lot over time
 - It's valuable to spend *some* time reading and thinking, but if this is all you do I think that's a mistake
- **Think from first principles** (sometimes)
 - Concrete exercise: Open a blank google doc, set a one hour timer, and start writing out your case for why AI Safety is the most important problem to work on. Spend the full hour on this, and if you run out of steam, go back through and poke at everything that feels confusing, or off, or dodgy. Write out all the counter-arguments you can think of, and repeat
 - This definitely isn't *all* you should do, but I think this is a really useful exercise for anything confusing!
- **Don't just try harder** - I have a failure mode I call [pushing the Try Harder button](#) where I label something as important and just try to channel a lot of willpower and urgency towards it. Don't do that! This takes a long time, and a lot will happen naturally as you think, talk to people, and do research.
 - If you find this really stressful, you have my permission to chill and not make it a priority for a while!
 - I've found my inside views develop a lot over time, fairly organically
- **Inside vs outside views is a spectrum** - there's no clear division between thinking for yourself and deferring. Forming inside views starts out by deferring, and then slowly forming more and more detailed models of where I'm deferring and why over time
 - My views have gone fairly organically from naive stories like 'AGI seems scary because intelligence is important and smart people think this matters' to more detailed ones like 'I think one reason AGI is scary is inner misalignment. Because neural networks have the base optimiser of stochastic gradient descent, the network may end up as a mesa-optimiser with a different mesa-objective. And this may create an instrumental

incentive for power seeking'. The latter story is way more detailed, but still includes a lot of implicit deferring - eg that we'll get AGI at all, that it'll be via deep learning, that mesa-optimisers are a thing at all, that there's an instrumental incentive for power seeking, etc. But expanding the tree of concepts like this is what progress looks like!

- Or, 'I should work on AI because AGI will happen eventually - if nature did it, so can we' to 'AGI is compute constrained. Using the bioanchors method to link to the size of the human brain gives 30-ish year AI timelines for human-level AI. I believe AGI is compute constrained because of some heuristic arguments about empirical trends, and because lots of smart people believe this'
- Getting here looks like downloading other people's gears level models into your head, and slowly combining them, deleting parts you disagree with, adding ideas of your own, etc

Misc

- **Defer intelligently** - Don't just adopt someone's opinions as your own because they're charismatic, high status, or well-credentialled. Think about *why* you think their opinions track the truth better than your own, and in which areas you're willing to defer to them. Figure out how hard they've thought about this, and whether they've taken the belief seriously
 - One key question is how much feedback they get from the world - would they know if they were wrong? I think some fields score much better on this than others - I'm a lot more comfortable disagreeing with many moral philosophy professors and being a committed consequentialist than I am with eg disagreeing with most algebraic geometers. Mathematicians get feedback re whether their proofs work in a way that, as far as I can tell, moral philosophy doesn't
 - And be domain specific - I'd defer to a Cambridge maths professor about mathematical facts, but not on a topic like 'how best to teach maths to undergraduates' - they clearly haven't done enough experimentation to tell if they're missing out on vastly better methods
- **You can act without an inside view**
 - Forming a good inside view takes a really long time! I've been doing full-time safety research for the past year and a bit and I'm still very confused
 - An analogy - a PhD is essentially a training program to give people an inside view for a specific research area. And this takes several years! IMO a question like 'is AGI an existential risk' is much harder than most thesis topics, and you don't have a hope of *really* understanding it without that much work
 - You can always change your mind and pivot later! Make the best expected value bet given what you know at the time, and what information you might get in future
 - Gathering information has costs! Sometimes thinking harder about a problem is analysis paralysis, and it's worth just running with your best guess
 - I think it's good to spend maybe 10% of your time long-term on high-level thinking, strategy, forming inside views, etc - a lot of your time should be spent actually doing stuff!
 - Though it's OK to spend a higher percentage early on when you have major decisions like what career path to go down.

- **You don't have to form an inside view** - Forming inside views that track the truth is *hard*, and it's a skill. You might just be bad at it, or find it too stressful. And this is fine! It shouldn't be low-status or guilt-inducing to just do what people more competent than you recommend
 - You can be a great research assistant, ops person, engineer etc without having a clear inside view - just find someone smart who you trust, explain your situation, and do what they think is best
 - I think the main reason this is a bad idea is motivational, not really about truth-tracking. And it's up to you how much you care about this motivationally!
 - An analogy: I think basically all AI Safety researchers who have ideas for an agenda should get funded, even if I personally think their agenda is BS. Likewise, I want them all to have enough labour available to execute well on their agenda - picking the agenda you're the best personal fit for and just deferring is a good way to implement this in practice.
- **Aim high, but be OK with missing** - It's valuable and important practice to try forming inside views, but it's also pretty hard! It's OK to struggle and not make much progress
 - IMO, trying to think for yourself is great training - it'll help you think more clearly, be harder to con, become a better researcher, etc.
 - Outside view: The vast majority of the world thinks AI Safety is nonsense, and puts very few resources towards it. This is worth taking seriously! You shouldn't throw your life away on a weird and controversial idea without thinking seriously about it first
 - This is a good way to trade-off between motivation and truth-tracking - so long as I try hard to think for myself, I feel OK motivationally, even if I know that I may not be tracking truth well
 - In practice, I try hard to form my own views, but then make big decisions by deferring a lot and forming an all-things-considered view, which I expect to track truth better
 - If you aren't doing full-time research, it's *much* harder to form clear views on things! This is a really hard thing you're trying to do
- **Convey mindsets, not inside views** - If you're talking to someone else about this stuff, eg while community building, it's important to try to convey the *spirit* and *mindset* of forming inside views, more so than your actual views. Try to convey all of the gears-level models in your head, but make it clear that they're just models! Try to convey what other people believe in.
 - I try hard to be clear about which beliefs I'm confident in, which are controversial, which points I'm deferring on, and which things I've thought hard about. I think this is important for avoiding information cascades, and building a healthy community
 - Relatedly, if you're mostly doing community building, it's totally fine to not have inside views on hard technical questions like AI Safety! Your goal is more to help people in your community form their own views on things - having views of your own is helpful but not essential.

Prediction Markets are for Outcomes Beyond Our Control

Betting markets are the gold standard of expert predictions because bets are the ultimate test of what people truly believe.

The best betting markets are highly liquid. A liquid market is one where you can place a large bet without moving the price very much. Liquid prediction markets work when no individual person can influence the outcome. Betting markets are a great way to find out if "it will rain tomorrow" or whether "candidate x will be elected president next year".

But what if a single person can influence the outcome? For example, what would happen if I created a betting market for "Lsusr will publish a blog post tomorrow"?

Suppose I am ambivalent about whether I will publish a blog post tomorrow. If the price of "Lsusr will publish a blog post tomorrow" drops below 1.00 then I will buy shares of "Lsusr will publish a blog post tomorrow" and then pocket a risk-free profit by posting a blog post tomorrow. If the price of "Lsusr will publish a blog post tomorrow" rises above 0.00 then I will buy shares of "Lsusr will not publish a blog post tomorrow" and then pocket a risk-free profit by not posting a blog post tomorrow.

The market equilibrium occurs even if I am unaware that the prediction market exists. Suppose the price drops to 0.99. A trader could buy shares and then pay me a small fee to influence the outcome.

I am not truly ambivalent. Suppose I'm willing to influence the outcome in exchange for \$500. What happens? If the market liquidity is less than \$500 then we have a functional prediction market. If the market liquidity is more than \$500 then we have a regular market.

Prediction markets function best when liquidity is high, but they break completely if the liquidity exceeds the price of influencing the outcome. Prediction markets function only in situations where outcomes are expensive to influence.

Simplify EA Pitches to "Holy Shit, X-Risk"

This is a linkpost for <https://www.neelnanda.io/45-x-risk>

TL;DR If you believe the key claims of "there is a $>=1\%$ chance of AI causing x-risk and $>=0.1\%$ chance of bio causing x-risk in my lifetime" this is enough to justify the core action relevant points of EA. This clearly matters under most reasonable moral views and the common discussion of longtermism, future generations and other details of moral philosophy in intro materials is an unnecessary distraction.

Thanks to Jemima Jones for accountability and encouragement. Partially inspired by Holden Karnofsky's excellent [Most Important Century](#) series.

Disclaimer: I recently started working for Anthropic, but this post entirely represents my opinions and not those of my employer

Introduction

I work full-time on AI Safety, with the main goal of reducing x-risk from AI. I think my work is really important, and expect this to represent the vast majority of my lifetime impact. I am also highly skeptical of total utilitarianism, vaguely sympathetic to person-affecting views, prioritise currently alive people somewhat above near future people and significantly above distant future people, and do not really identify as a longtermist. Despite these major disagreements with some common moral views in EA, which are often invoked to justify key longtermist conclusions, I think there are basically no important implications for my actions.

Many people in EA really enjoy philosophical discussions and debates. This makes a lot of sense! What else would you expect from a movement founded by moral philosophy academics? I've enjoyed some of these discussions myself. But I often see important and controversial beliefs in moral philosophy thrown around in introductory EA material (introductory pitches and intro fellowships especially), like strong longtermism, the astronomical waste argument, valuing future people equally to currently existing people, etc. And I think this is unnecessary and should be done less often, and makes these introductions significantly less effective.

I think two sufficient claims for most key EA conclusions are "AI has a $>=1\%$ chance of causing human extinction within my lifetime" and "biorisk has a $>=0.1\%$ chance of causing human extinction within my lifetime". I believe both of these claims, and think that you need to justify at least one of them for most EA pitches to go through, and to try convincing someone to spend their career working on AI or bio. **These are really weird claims.** The world is clearly not a place where most smart people believe these! If you are new to EA ideas and hear an idea like this, with implications that could transform your life path, **it is right and correct to be skeptical.** And when you're making a complex and weird argument, it is *really* important to distill your case down to the minimum possible series of claims - each additional point is a new point of inferential distance, and a new point where you could lose people.

My ideal version of an EA intro fellowship, or an EA pitch (a $>=10$ minute conversation with an interested and engaged partner) is to introduce these claims and a minimum viable case for them, some surrounding key insights of EA and the mindset of doing good, and then digging into them and the points where the other person doesn't agree or feels confused/skeptical. I'd be excited to see someone make a fellowship like this!

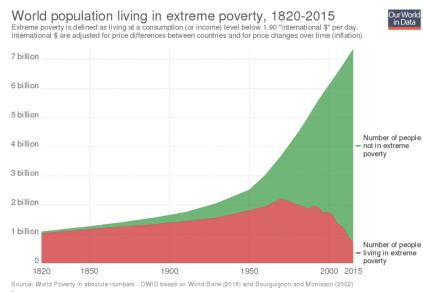
My Version of the Minimum Viable Case

The following is a rough outline of how I'd make the minimum viable case to someone smart and engaged but new to EA - this is intended to give inspiration and intuitions, and is something I'd give to open a conversation/Q&A, but is not intended to be an airtight case on its own!

Motivation

Here are some of my favourite examples of major ways the world was improved:

- Norman Borlaug's Green Revolution - One plant scientist's study of breeding high-yield dwarf wheat, which changed the world, converted India and Pakistan from grain importers to grain exporters, and likely saved over 250 million lives
- The eradication of smallpox - An incredibly ambitious and unprecedented feat of global coordination and competent public health efforts, which eradicated a disease that has killed over 500 million people in human history
- Stanislav Petrov choosing *not* to start a nuclear war when he saw the Soviet early warning system (falsely) reporting a US attack
- The industrial and scientific revolutions of the last few hundred years, which are responsible for [this incredible graph](#).



When I look at these and other examples, a few lessons become clear if I want to be someone who can achieve massive amounts of good:

- Be willing to be ambitious
- Be willing to believe and do weird things. If I can find an important idea that most people don't believe, and can commit and take the idea seriously, I can achieve a lot.
 - If it's obvious, common knowledge, someone else has likely already done it!
 - Though, on the flipside, *most* weird ideas are wrong - don't open your mind so much that your brains fall out.
- Look for **high-leverage!**

- The world is big and inter-connected. If you want to have a massive impact, it needs to be leveraged with something powerful - an idea, a new technology, exponential growth, etc.

When I look at today's world through this lens, I'm essentially searching for things that could become a really big deal. Most things that have been really big, world-changing deals in the past have been some kind of major emerging technology, unlocking new capabilities and new risks. Agriculture, computers, nuclear weapons, fossil fuels, electricity, etc. And when I look for technologies emerging *now*, still in their infancy but with a lot of potential, AI and synthetic biology stand well above the rest.

Note that these arguments work about as well for focusing on highly leveraged positive outcomes or negative outcomes. I think that, in fact, given my knowledge of AI and bio, that there are plausible negative outcomes, and that reducing the likelihood of these is tractable and more important than ensuring positive outcomes. But I'd be sympathetic to arguments to the contrary.

AI - 'AI has a $\geq 1\%$ chance of x-risk within my lifetime'

The human brain is a natural example of a generally intelligent system. Evolution produced this, despite a bunch of major constraints like biological energy being super expensive, needing to fit through birth canals, using an extremely inefficient optimisation algorithm, and intelligence not obviously increasing reproductive fitness. While evolution had the major advantage of four billion years to work with, it seems highly plausible to me that humanity can do better. And, further, there's no reason that human intelligence should be a limit on the capabilities of a digital intelligence.

On the outside view, this is **incredibly important**. We're contemplating the creation of a second intelligence species! That seems like one of the most important parts of the trajectory of human civilisation - on par with the dawn of humanity, the invention of agriculture and the Industrial Revolution. And it seems crucial to ensure this goes well, especially if these systems end up much smarter than us. It seems plausible that the default fate of a less intelligent species is that of gorillas - humanity doesn't really bear gorillas active malice, but they essentially only survive because we want them to.

Further, there are specific reasons to think that this could be really scary! AI systems mostly look like optimisation processes, which can find creative and unexpected ways to achieve these objectives. And [specifying the right objective is a notoriously hard problem](#). And there are good reasons to believe that such a system might have an instrumental incentive to seek power and compete with humanity, especially if it has the following three properties:

- **Advanced capabilities** - it has superhuman capabilities on at least some kinds of important and difficult tasks
- **Agentic planning** - it is capable of making and executing plans to achieve objectives, based on models of the world
- **Strategic awareness** - it can competently reason about the effects of gaining and maintaining power over humans and the real world

See [Joseph Carlsmith's excellent report](#) for a much more rigorous analysis of this question. I think it is by no means *obvious* that this argument holds, but I find it sufficiently plausible that we create a superhuman intelligence which is incentivised to seek power and successfully executes on this in a manner that causes human

extinction that I'm happy to put at least a 1% chance of AI causing human extinction (my fair value is probably 10-20%, with high uncertainty).

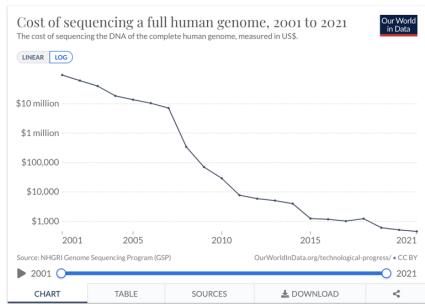
Finally, there's the question of timelines. Personally, I think there's a good chance that something like deep learning language models scale to human-level intelligence and beyond (and this is a key motivation of my current research). I find the [bio-anchors and scaling based methods of timelines](#) pretty convincing as an upper bound of timelines that's well within my lifetime. But even if deep learning is a fad, the field of AI has existed for less than 70 years! And it takes 10-30 years to go through a paradigm. It seems highly plausible that we produce human-level AI with some other paradigm within my lifetime (though reducing risk from an unknown future paradigm of AI does seem much less tractable)

Bio - 'Biorisk has a $>=0.1\%$ chance of x-risk within my lifetime'

I hope this claim seems a lot more reasonable now than it did in 2019! While COVID was nowhere near an x-risk, it has clearly been one of the worst global disasters I've ever lived through, and the world was highly unprepared and bungled a lot of aspects of the response. 15 million people have died, many more were hospitalised, millions of people have long-term debilitating conditions, and almost everyone's lives were highly disrupted for two years.

And things could have been much, much worse! Just looking at natural pandemics, imagine COVID with the lethality of smallpox (30%). Or COVID with the age profile of the Spanish Flu (most lethal in young, healthy adults, because it turns the body's immune system against itself).

And things get much scarier when we consider synthetic biology. We live in a world where multiple labs work on gain of function research, doing crazy things like trying to breed Avian Flu (30% mortality) that's human-to-human transmissible, and not all DNA synthesis companies will stop you trying to print smallpox viruses. Regardless of whether COVID was *actually* a lab leak, it seems at least plausible that it could have come from gain-of-function research on coronaviruses. And these are comparatively low-tech methods. Progress in synthetic biology happens fast!



It is highly plausible to me that, whether by accident, terrorism, or an act of war, that someone produces an engineered pathogen capable of creating a pandemic far worse than anything natural. It's unclear that this could actually cause human extinction, but it's plausible that something scary enough and well-deployed enough with a long incubation period could. And it's plausible to me that something which kills 99% of people (a much lower bar) could lead to human extinction. Biorisk is not my field and

I've thought about this much less than AI, but 0.1% within my lifetime seems like a reasonable lower bound given these arguments.

Caveats

- These are really weird beliefs! It is correct and healthy for people to be skeptical when they first encounter them.
 - Though, in my opinion, the arguments are strong enough and implications important enough that it's unreasonable to dismiss them without at least a few hours of carefully reading through arguments and trying to figure out what you believe and why.
 - Further, if you disagree with them, then the moral claims I'm dismissing around strong longtermism etc may be much more important. But you should disagree with the vast majority of how the EA movement is allocating resources!
- There's a much stronger case for something that kills *almost* all people, or which causes the not-necessarily-permanent collapse of civilisation, than something which kills *literally* everyone. This is a really high bar! Human extinction means killing everyone, including Australian farmers, people in nuclear submarines and bunkers, and people in space.
 - If you're a longtermist then this distinction matters a lot, but I personally don't care as much. The collapse of human civilisation seems super bad to me! And averting this seems like a worthy goal for my life.
 - I have an easier time seeing how AI causes extinction than bio
- There's an implicit claim in here that it's reasonable to invest a large amount of your resources into averting risks of extremely bad outcomes, even though we may turn out to live in a world where all that effort was unnecessary. I think this is correct to care about, but that this is a reasonable thing to disagree with!
 - This is related to the idea that we should maximise expected utility, but IMO importantly weaker. Even if you disagree with the formalisation of maximising expected value, you likely still agree that it's extremely important to ensure that bridges and planes have safety records far better than 0.1%
 - But also, we're dealing with probabilities that are small but not infinitesimal. This saves us from objections like Pascal's Mugging - [a 1% chance of AI x-risk is not a Pascal's Mugging](#).
 - It is also reasonable to buy these arguments intellectually, but not to feel emotionally able to motivate yourself to spend your life reducing tail risks. This stuff is hard, and can be depressing and emotionally heavy!
 - Personally, I find it easier to get my motivation from other sources, like intellectual satisfaction and social proof. A big reason I like spending time around EAs is that this makes AI Safety work feel much more viscerally motivating to me, and high-status!
- It's reasonable to agree with these arguments, but consider something else an even *bigger* problem! While I'd personally disagree, any of the following seem like justifiable positions: climate change, progress studies, global poverty, factory farming.
- A bunch of people do identify as EAs, but would disagree with these claims and with prioritising AI and bio x-risk. To those people, sorry! I'm aiming this post at the significant parts of the EA movement (many EA community builders, CEA, 80K, OpenPhil, etc) who seem to put major resources into AI and bio x-risk reduction

- This argument has the flaw of potentially conveying the beliefs of ‘reduce AI and bio x-risk’ without conveying the underlying generators of cause neutrality and carefully searching for the best ways of doing good. Plausibly, similar arguments could have been made in early EA to make a “let’s fight global poverty” movement that never embraced longtermism. Maybe a movement based around the narrative I present would miss the next Cause X and fail to pivot when it should, or otherwise have poor epistemic health.
 - I think this is a valid concern! But I also think that the arguments for “holy shit, AI and bio risk seem like really big deals that the world is majorly missing the ball on” are pretty reasonable, and I’m happy to make this trade-off. “Go work on reducing AI and bio x-risk” are things I would love to signal boost!
 - But I have been deliberate to emphasise that I am talking about *intro* materials here. My ideal pipeline into the EA movement would still emphasise good epistemics, cause prioritisation and cause neutrality, thinking for yourself, etc. But I would put front and center the belief that AI and bio x-risk are substantial and that reducing them is the biggest priority, and encourage people to think hard and form their own beliefs
- An alternate framing of the AI case is “Holy shit, AI seems really important” and thus a key priority for altruists is to ensure that it goes well.
 - This seems plausible to me - it seems like the downside of AI going wrong could be human extinction, but that the upside of AI going really well could be a vastly, vastly better future for humanity.
 - There are also a lot of ways this could lead to bad outcomes beyond the standard alignment failure example! [Maybe coordination just becomes much harder in a fast-paced world of AI](#) and this leads to war, or we pollute ourselves to death. Maybe it massively accelerates technological progress and we discover a technology more dangerous than nukes and with a worse Nash equilibria and don’t solve the coordination problem in time.
 - I find it harder to imagine these alternate scenarios literally leading to extinction, but they might be more plausible and still super bad!
 - There are some alternate pretty strong arguments for this framing. One I find very compelling is drawing an analogy between exponential growth in the compute used to train ML models, and the [exponential growth in the number of transistors per chip](#) of Moore’s Law.
 - Expanding upon this, [historically most AI progress has been driven by increasing amounts of computing power and simple algorithms that leverage them](#). And [the amount of compute used in AI systems is growing exponentially](#) (doubling every 3.4 months - compared to Moore’s Law’s 2 years!). Though the rate of doubling is likely to slow down - it’s much easier to increase the amount of money spent on compute when you’re spending less than the millions spent on payroll for top AI researchers than when you reach the order of magnitude of figures like Google’s \$26bn annual R&D - it also seems highly unlikely to stop completely.
 - Under this framing, working on AI now is analogous to working with computers in the 90s. Though it may have been hard to predict exactly how computers would change the world, there is no question that they did, and it seems likely that an ambitious altruist could have gained significant influence over how this went and nudged it to be better.
 - I also find this framing pretty motivating - even if specific stories I’m concerned by around eg inner alignment are wrong, I can still be pretty confident that *something* important is happening in AI, and my

research likely puts me in a good place to influence this for the better.

- I work on [interpretability research](#), and these kind of robustness arguments are one of the reasons I find this particularly motivating!

Abstractions as Redundant Information

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

What is it that makes the concept of “pencil” a good abstraction?

One way to frame it: there are lots of “pencils” in the world - lots of little chunks of the world which I could look at which all contain information about pencils-in-general. Many different places from which I could gain information about “pencils”, and many different places where I can apply that information to make predictions. If I don’t know that pencils are usually made of wood with a graphite core, there are many different chunks of the world which I could look at to gain that information. If I do know that pencils are usually made of wood with a graphite core, there are many different chunks of the world where I could apply that information to predict the materials from which something is made.

Alternatively, consider a gear, spinning in a gearbox. What makes the gear’s rotational speed a good abstraction? Well, there are lots of little patches of metal comprising the gear. If I don’t know the gear’s rotation speed, I could precisely estimate it from any of the little patches. If I do know the gear’s rotation speed, I can use it to predict the rotation of many different little patches of metal within the gear.

This seems like an intuitively-reasonable notion of what makes abstractions “good” in general: there are many different places from which we can learn the information, and many different places where we can apply the information to make predictions. In other words, a good abstraction is highly redundant: it appears in many different places, and in any of those places we can use the abstract information to make predictions and/or gain more information about the abstraction in general.

In this post I’ll sketch out one way to formalize this idea mathematically, and show that it’s equivalent to the formalization of [abstraction as information-at-a-distance](#). In particular, in the gear example: **conditional on the highly redundant information (i.e. the overall rotation of the gear), the low-level rattling of far-apart chunks of metal is statistically independent**. More generally, redundancy yields the same high-level abstract information as the [Telephone Theorem](#), but in a mathematically-cleaner form, and without the need for different summaries between different variables.

Meta

Advice for readers: I try to keep the more-dense math in a few specific sections and the appendices, which can all be skimmed/skipped if you just want a conceptual picture.

Epistemic status: I’m aiming for physics-level rigor here. The proofs involve some shenanigans with infinite limits, and I expect them to contain subtle errors. However, I also expect that the results will accurately predict reality in practice, and that whatever subtle errors the proofs contain can be patched by the sort of mathematician who enjoys dealing with tricky limit shenanigans.

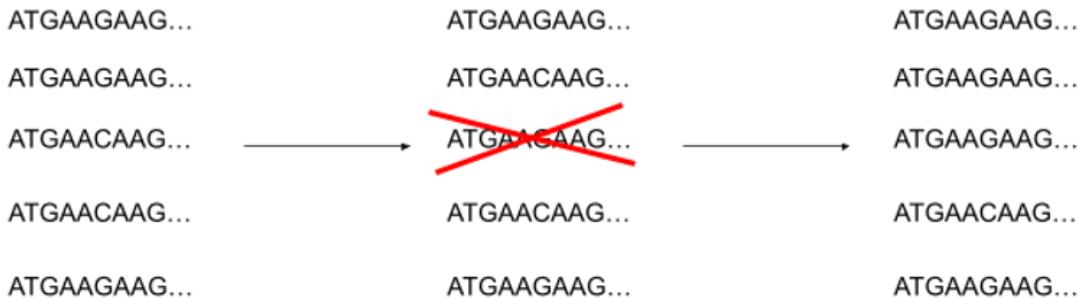
Thankyou to Rob Miles and Adam Shimi for suggesting I try [Excalidraw](#).

Basic Idea: Redundant Information Is Conserved Under Resampling

Suppose I sample the genomes of two random humans, G_1 and G_2 . What information is redundant across these two random variables?

One intuitively-reasonable answer: if I threw away one of the two sequences, then the redundant information is whatever I could still figure out from the other. More generally, if I have a whole bunch of genomes from random humans, $G_1 \dots G_N$, the redundant information is whatever I could still figure out after throwing one of them away.

To formalize “what I could still figure out after throwing one away”, we’ll use the idea of *resampling*: I throw away the value of G_i , and then sample a *new* value for G_i , using my original probability distribution *conditioned on* all the other genomes. So, for instance, I throw away G_i , then I look at all the other genomes and see that in most places they’re the same - so when I sample my new G_i , I know that it should match all the other genomes in all those places. However, there are a few locations where the other genomes differ, e.g. maybe 10% of them contain a particular mutation. So, when I sample my new G_i , it will contain that mutation with roughly 10% probability (assuming there’s enough data to swamp the impact of my priors).



Resampling: we throw out one genome, then resample it from a distribution informed by all the other genomes. Any information which is highly redundant across the genomes is conserved - e.g. the sequence prefix “ATGAA” stays the same.

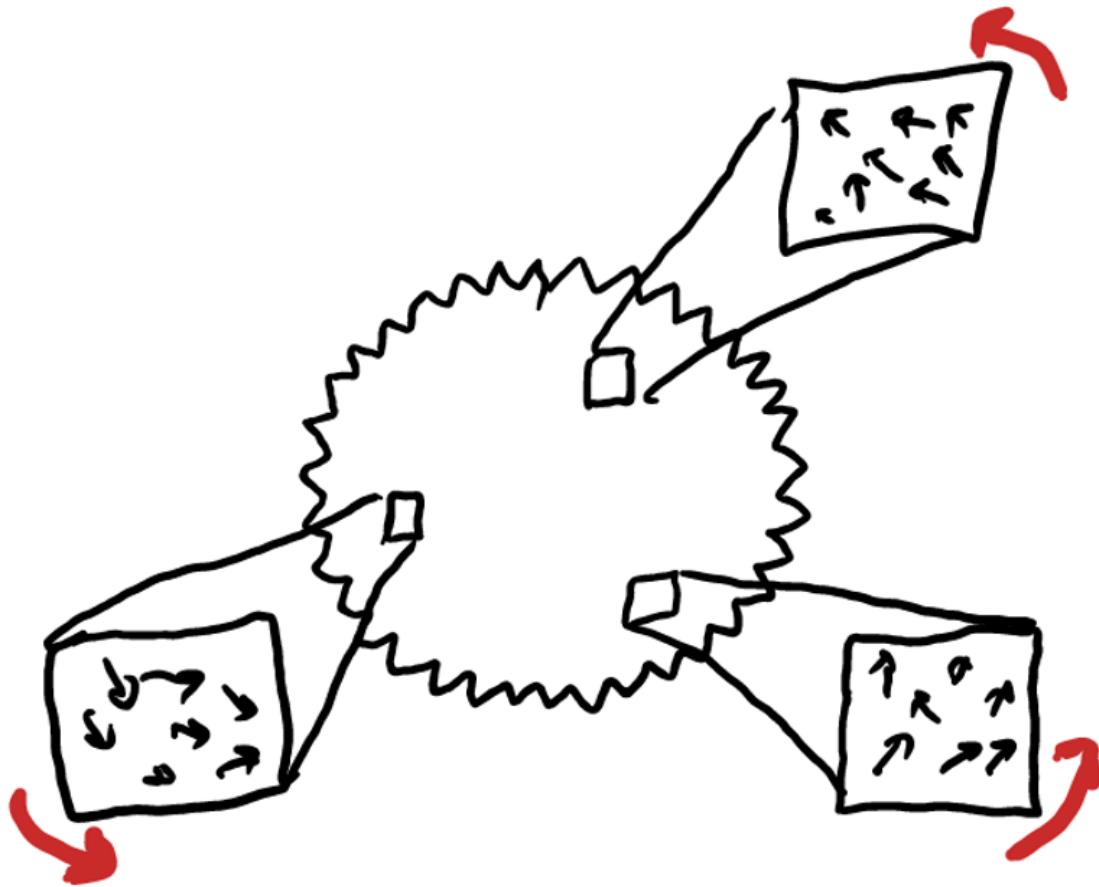
Now I repeat this process many times, each time throwing away one randomly-chosen genome and resampling it conditional on the others. Intuitively, I expect that the information conserved by this process will be whatever information is highly redundant, so that approximately-zero loss occurs at each step.

General method:

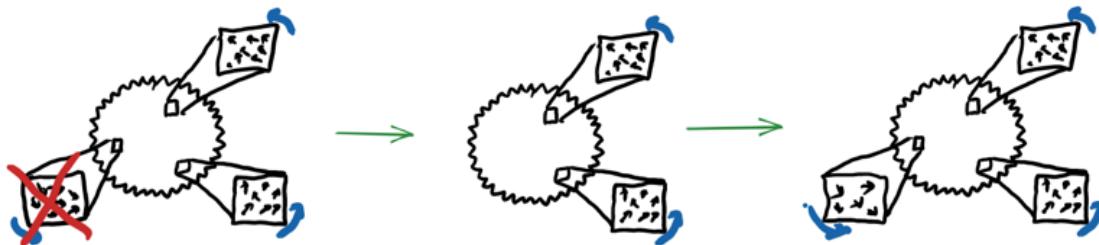
- Start with a bunch of random variables $X_1 \dots X_N$, a joint distribution $P[X|M_{\text{base}}]$, and a value X_i^0 for each variable. (See the Equations section below for more details on the notation.)
- At each “timestep”, pick one variable at random, throw away its value, and resample it conditional on all the other variable values.
- Run this process for a long time.
- See what information about the initial conditions X^0 is conserved.

Conceptual Example: Gear

Suppose I have a gear, spinning around in a gearbox. I look at a few nanometer-size patches on the gear’s surface, just a hundred-ish atoms each, and measure the (approximate) position and velocity of each atom in each little patch. Notation: random variable X_i gives the positions and velocities of each atom in patch i .



What information is redundant across these different patches? Intuitively, I can look at any patch and average together the rotational velocities of the atoms about the gear's center to get a reasonably-precise estimate of the gear's overall rotation speed. If I throw away X_i , I can still precisely estimate the gear's overall rotation from the other X 's. Then, when I resample X_i , I will resample it so that the average rotational velocity of the atoms in X_i matches the gear's overall rotation.



Equations

Notation: we'll use X^t for the variable-values after t steps of this process, and X^∞ for variable-values after running the process for an arbitrarily long time. So, we're mainly interested in the mutual information between X^0 and X^∞ . The resampling process itself specifies the distribution $P[X^\infty | X^0, M_{\text{resample}}]$.

We'll use "model" variables M_{base} and M_{resample} to distinguish probabilities from the "base" distribution (which is only over $X_1 \dots X_N$) vs probabilities from the resampling process (which is

over $X_1 \dots X_N, X_1 \dots X_N, \dots, X_1 \dots X_N$). One potential point of confusion: both models contain a variable called "X", but these two variables are "in different scopes" (in the programmer's sense); "X" means something different depending on which model it's in. In the base model, X is just a single instance of our base random variables. In the resampling model, X consists of many instances X^t , one for each timestep t, and each individual instance is distributed the same as the base model's X. We use the same variable name for both because there's a conceptual correspondence between the two.

The resampling process defines the full relationship between the two models:

$$P[X_1 = x_1 \dots X_N = x_N | M_{\text{resample}}] = P[X_1 = x_1 \dots X_N = x_N | M_{\text{base}}]$$

$P[X_i = x_i | M_{\text{resample}}, X^{t-1} = x] = P[X_i = x_i | M_{\text{base}}, X_{\neq i} = x_{\neq i}]$ (assuming variable i is resampled at resampling-time t; for all the other variables, $P[X_i = x_i | M_{\text{resample}}, X^{t-1} = x] = I[x_i = x_i]$, since their values just stay the same)

$$P[X | M_{\text{resample}}] = P[X^0 | M_{\text{resample}}] \prod_t P[X^t | M_{\text{resample}}, X^{t-1}]$$

These equations follow directly from the process outlined above, and define the distribution $P[X|M_{\text{resample}}]$ in terms of the distribution $P[X|M_{\text{base}}]$.

... So We're Running MCMC?

If you've worked with Markov Chain Monte Carlo (MCMC) algorithms before, this should look familiar: we're basically asking which information about the initial conditions will be conserved as we run a standard MCMC process.

If you've worked with MCMC algorithms before, you might also guess that this question has two answers:

- In theory, assuming $P[X|M_{\text{base}}] > 0$ everywhere, the resampling-distribution always approaches $P[X|M_{\text{base}}]$ as we run the process regardless of initial conditions. Since we always approach the same distribution regardless of initial conditions, no information about the initial conditions is conserved.
- In practice, the time it takes for that limit to kick in increases (often exponentially) with the number of variables in the model, so lots of information is conserved in large models over any practically-achievable number of steps of the process.

In this post, we're going to think about infinitely large models, so the process no longer converges to $P[X|M_{\text{base}}]$ at all; that convergence time goes to infinity. The distribution does still converge, but the limiting distribution depends on the initial conditions, and that dependence is

exactly what we're interested in. Unfortunately, this introduces a bunch of tricky subtleties about how we take the limits: do we take the limit of infinitely many variables first, or the limit of infinite time first, or do we take them at the same time with some fixed relationship between the two? I'll handle those subtleties mainly by ignoring them and hoping a mathematician comes along to clean it up later. Remember, we're aiming for physics-level rigor here.

The important takeaway of this section is that we have a ton of data on how these sorts of processes actually behave in practice, thanks to the popularity of MCMC algorithms. So we don't just have to rely on physics-level-of-rigor arguments; anyone with firsthand experience with MCMC on large models can use their intuition as a guide. (I mostly won't explicitly talk more here about lessons from MCMC; I expect that those of you already familiar with the topic can reason it through for yourselves, and explaining the relevant experience/intuitions to people not already familiar is beyond the scope of this post.)

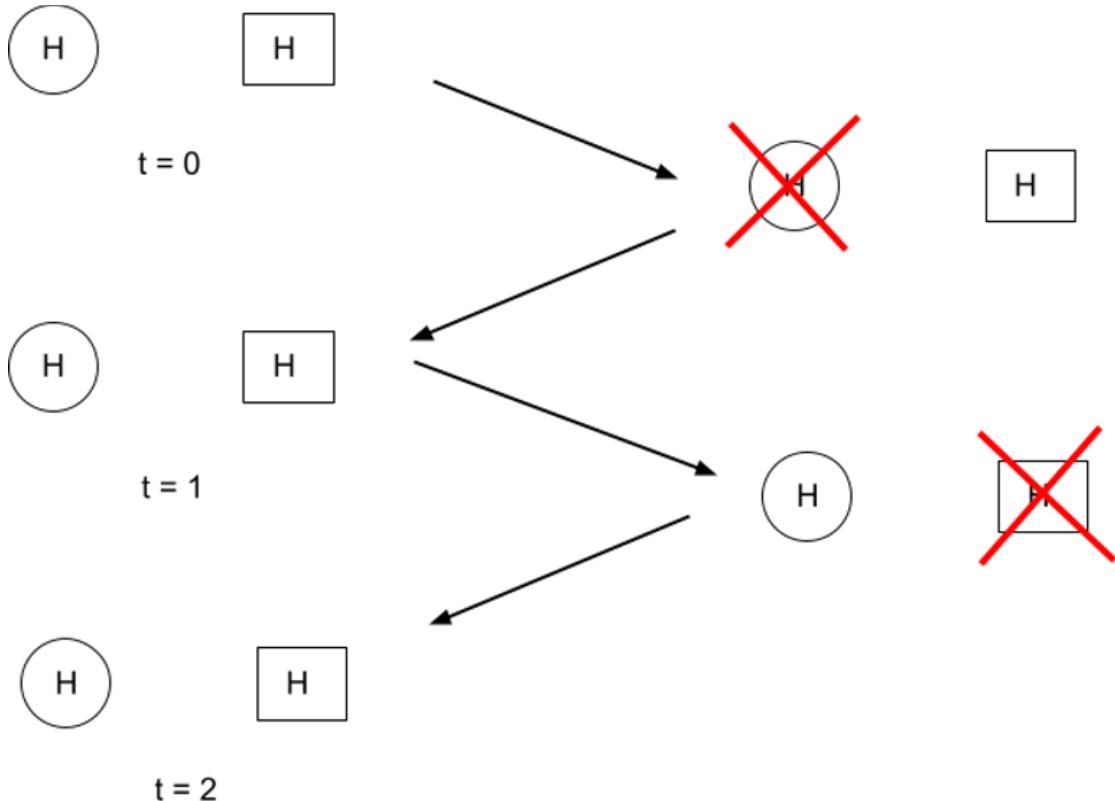
Worked Examples

Two Variables

Let's start with a trivial example to show how any information at all can be conserved by the resampling process. We have two random variables, X_1 and X_2 . Our model for the two variable values is:

- X_1 is an unbiased coin flip
- X_2 is a record on a piece of paper indicating whether the coin came up heads or tails ("H" if the coin came up heads, "T" if the coin came up tails)

What happens when we run our resampling process on this system? Well, first we throw away the coin flip, and resample it given our record. If the record says "H", then we know the coin came up heads, so our sampler selects heads again for X_1 ; vice-versa for tails. The first coin is therefore reset to its original value. Then, we throw away the record, and resample it given our coin. If the coin is heads, we know the record says "H"; vice-versa for tails. The record is therefore reset to its original value.



Yes, I know, it's poor taste mixing image styles. But this post has already been in the pipe for weeks, and I have other upcoming posts which need to cite it, so it's time to cut corners.

The process continues, back and forth, with each variable "storing the information" when the other is thrown away, and the information then perfectly copied back over into the resampled variable value.

On the other hand, imagine that our record-keeping is imperfect - maybe there's a 10% chance that X_2 records the wrong value. Then, at each "timestep" of the resampling process, there's roughly a 20% chance (10% for each variable) that we'll lose the original value. Given, say, 100 timesteps, we'll lose approximately-all of the information about the original values.

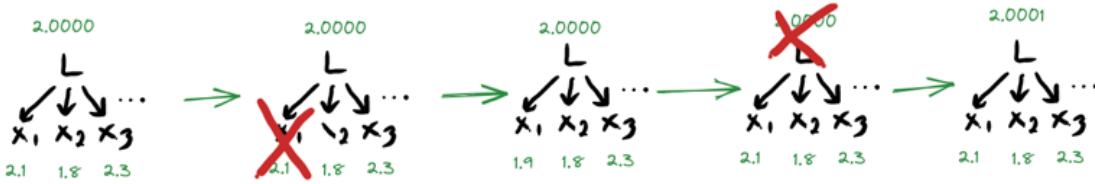
General point: only information which is perfectly conserved at each timestep will be conserved indefinitely; everything else is completely lost.

In general, the information perfectly conserved indefinitely will be the value of deterministic constraints, i.e. functions f_1 and f_2 such that $f_1(X_1) = f_2(X_2)$ with probability 1 in the base distribution $P[X|M_{\text{base}}]$. (We can prove this via the Telephone Theorem plus an equilibrium condition, but it's not the main theorem of interest in this post.)

Many Measurements Of One Thing

Let's say we have a stick of length L , and take N conditionally-independent measurements $X_1 \dots X_N$ of L . We'll model each measurement as normally distributed with mean L and standard

deviation σ , and we'll assume that we have enough data that the prior on L doesn't matter much (i.e. we'll just pretend the prior is flat).



To simplify the analysis a little, we'll resample the variables in order rather than at random - i.e. we resample L conditional on all the measurements, then resample each of the measurements $X_1 \dots X_N$ conditional on L , and repeat.

When we resample L , we draw from a normal distribution with mean equal to the average of the measurements (i.e. $\frac{1}{N} \sum_i X_i$) and standard deviation $\frac{\sigma}{\sqrt{N}}$, so our new L will be about $\frac{\sigma}{\sqrt{N}}$ away from the previous average. When we resample the measurements, their new average will be normally distributed with mean L and standard deviation $\frac{\sigma}{\sqrt{N}}$, so the new average will be about $\frac{\sigma}{\sqrt{N}}$ away from the previous L . In other words: L and the measurement average follow a random walk, drifting about $\frac{\sigma}{\sqrt{N}}$ per timestep. Over T timesteps, they will drift a distance of about $\sqrt{T} \frac{\sigma}{\sqrt{N}}$.

(In general, the distance a random walk drifts scales with \sqrt{T} rather than T , since it often wanders back on itself.)

So: if we run the process for a number of steps $T \gg N$, then all information about the initial conditions is lost. On the other hand, if the number of variables $N \gg T$, then the drift is close to zero, so L and the measurement average are approximately conserved. The order in which we take our limits matters.

Practically speaking, we're looking for information which is *approximately* conserved, i.e. the "timescale" T over which it's lost is large. So it makes sense to consider L and the measurement average as approximately-conserved when N is large. That's our abstract information in this example.

Factorization

Now for our first theorem about this kind of resampling process.

Imagine that our base distribution is *local* - i.e. each variable only "directly interacts with" a few "neighbor variables". When modeling a physical gear, for instance, each little chunk of metal only interacts directly with the chunks of metal spatially adjacent. Any longer-range interactions have to "go through" those direct interactions.

Our theorem says that interactions are still local after controlling for the information conserved by the resampling process. In the gear example, after controlling for the high-level rotation of the gear, the remaining low-level vibrations and rattling are still local; the low-level details of each chunk of metal interact directly only with the low-level details in chunks spatially adjacent.

Why does this matter? In general, locality is the main tool which [lets us reason about our high-dimensional world at all](#). It means we can look at one part of the world, and understand what's going on there without having to understand everything that's happening in the whole universe. The factorization theorem says that this still applies when we condition on our high-level knowledge - in other words, we can "zoom in" on lower-level details, and add them to our high-level picture without having to understand all the other low-level details in the whole universe. Conditional on our high-level knowledge, any low-level information still has to flow through neighboring variables in order to influence things "far away" in the graph.

That's going to be key to our next theorem, which is the main item of interest in this post.

Formal Statement

Resampler Conserves Locality: If the base distribution $P[X|M_{\text{base}}]$ factors over some graph G , then so does the limiting resampling distribution $P[X^\infty|M_{\text{resample}}, X^0]$. This factorization theorem applies to both undirected graphical models (i.e. Markov Random Fields) and directed graphical models (i.e. Bayes Nets/Causal Models). See the appendices for a proof sketch.

In the gear example, the graph G would be the adjacency graph for chunks of metal: each chunk is a node, and the edges show spatial adjacency. The factorization follows the standard formulas for factorization of Markov Random Fields or Bayes Nets, depending on which type of graphical model we're using.

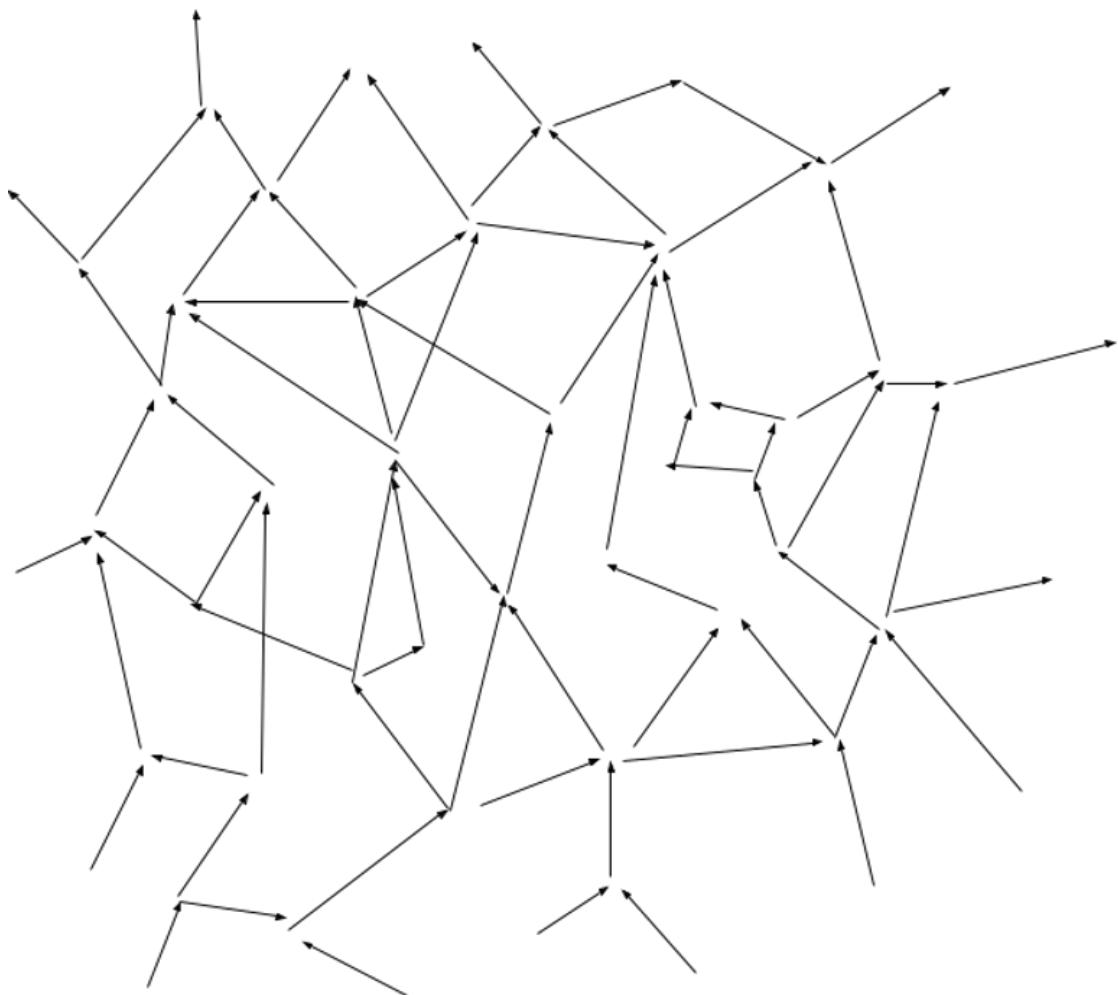
The Interesting Part: Resampler-Conserved Quantities Mediate Information At A Distance

Time for the big claim from earlier: in the gear example, conditional on the highly redundant information (i.e. the overall rotation of the gear), the low-level rattling of far-apart chunks of metal is statistically independent.

More generally: assume that our base distribution factors on a graph G . **Conditional on all the quantities perfectly conserved by the resampling process, variables far apart in G are independent.** If you've read the [Telephone Theorem](#), this is basically the same, but with one big upgrade: our "high-level summary" no longer depends on *which* notion of "far away" we use; the *same* summary applies to *any* sequence of nonoverlapping nested Markov blankets. We can take the information conserved by the resampling process to be the "high-level abstractions" for the whole model.

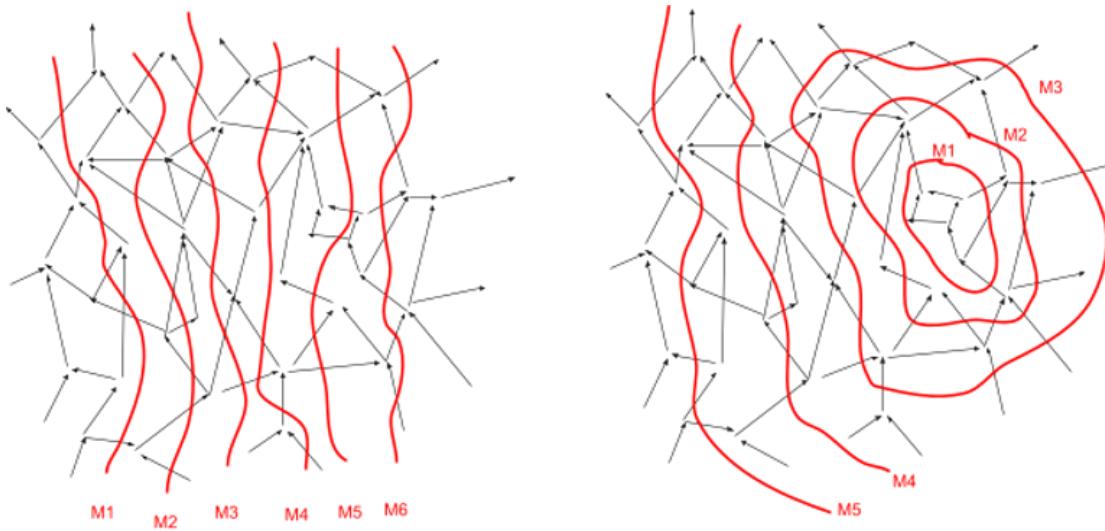
Let's unpack that. First, we'll do a quick recap of the Telephone Theorem.

We start with a large Causal Model/Bayes Net:



Each node is a random variable, and the arrows show direct causal influence; paths show indirect causal influence. If you've used MCMC before, you were probably picturing something like this already; we usually use MCMC with Bayes nets, because the locality structure makes it easy to resample variables (we only need to look at neighbor values rather than the values of all variables).

Now, just like in the Telephone Theorem, we picture a sequence of nested Markov blankets $M_1 \dots M_n$ in our model:



Each possible sequence of blankets cuts the graph into pieces, with each piece only connected directly to the piece before and the piece after. A choice of sequence of blankets defines a notion of “far away” - i.e. if two variables are separated by a large number of “layers” of blankets in the sequence, then they are “far apart”. In order for M_1 to have any mutual information with M_n , that information must propagate through each of the layers in between.

The basic idea of the Telephone Theorem is that information is either perfectly conserved or completely lost as we move through enough layers; information can only propagate “far away” if the information can be perfectly computed from each layer individually.

... but if some information can be perfectly computed from each layer individually, then that information will be conserved by our resampler. When resampling M_n , I can still perfectly calculate the information from M_{n+1} or M_{n-1} , and therefore the information will be perfectly conserved. So, the only information which can propagate far away is information which is perfectly conserved by resampling. That means that *conditional* on the information conserved by resampling, the mutual information must drop to zero.

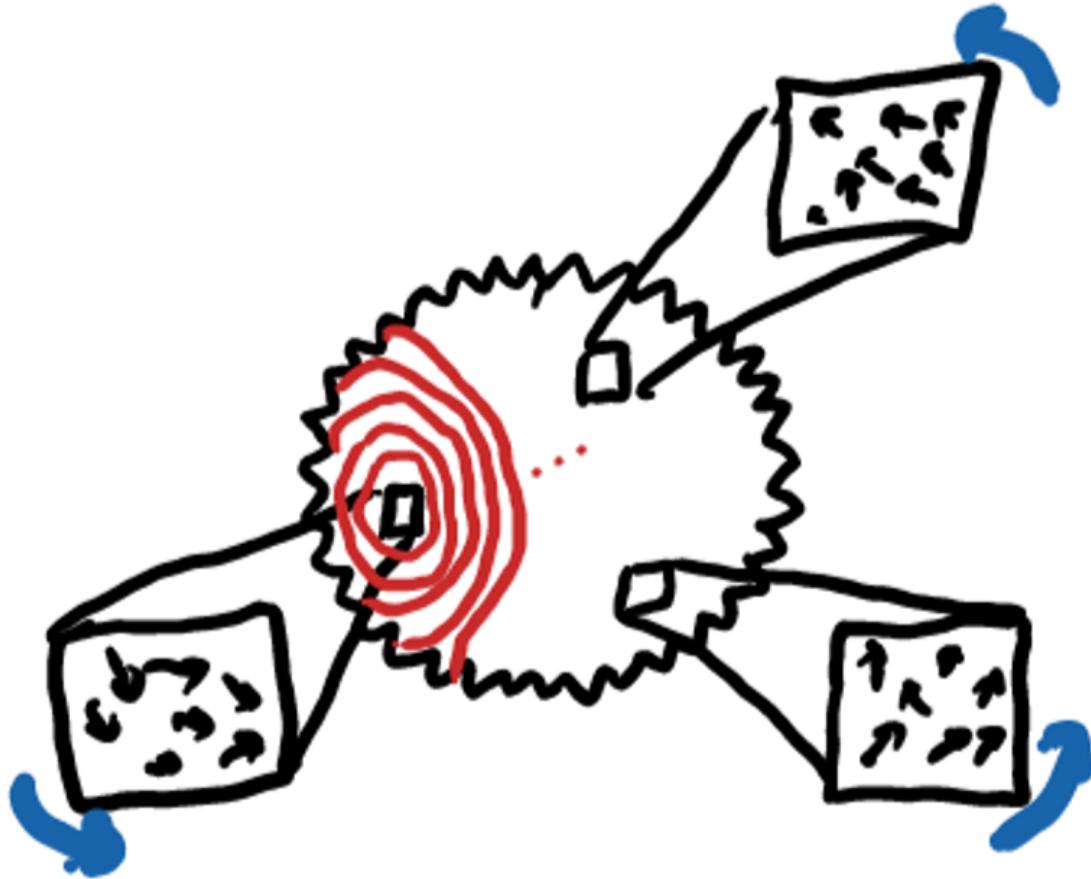
A slightly more formal version of that argument is in the appendices, but that’s the core idea.

Formal Statement

Let F satisfy $P[X^\infty | M_{\text{resample}}, X^0] = P[X^\infty | M_{\text{resample}}, F(X^0)]$, i.e. F encodes the values of all conserved quantities in the resampling process. For any infinite sequence of nested nonoverlapping Markov blankets B_1, B_2, \dots on the base model, the conditional mutual information $M I(B_1, B_n | F(X)) \rightarrow 0$ as $n \rightarrow \infty$.

Gear Example

In the gear example, our Markov blankets might be nested layers of metal:



"Far away" then indicates moving through a large number of such layers.

In order for information to propagate far away, we must be able to compute it from each layer - i.e. we can very precisely compute the overall rotation of the gear from the overall rotation of chunks of metal in each layer. So, when we resample a chunk of metal, the overall rotation will be conserved - it will be "stored in" the other chunks.

This reasoning must apply to any information which propagates through many layers: if the information propagates far away, it will be conserved by resampling. So, assuming the overall rotation is the only quantity conserved by the resampling process, far-apart chunks of the gear must be independent given the overall rotation.

Intuitively, this lets us factor the gear into a "global" component and a "local" component. The global component, the gear's overall rotation, is redundantly represented; it can be estimated by looking at many different little chunks, and we expect these estimates to (approximately) agree. The local component captures everything else, and is guaranteed to be "local" in the sense that far apart pieces are independent.

Conclusion

We started with the intuition that abstraction is about redundant information: there are many different places from which we can learn the information, and many different places where we can apply the information to make predictions. That's what makes abstractions generalizable and useful.

Then, we showed that a formalization of this intuition based on resampling variables reproduces the main ideas of abstraction as information-relevant-at-a-distance. In particular, the resampling

approach yields a better version of the Telephone Theorem.

Personally, I came to all this in a different order: I noticed that the Telephone Theorem required redundancy of information, figured out the resampling thing, and only backed out the intuition of abstraction-as-redundant-information later. Nonetheless, it was an exciting thing to find: when different intuitive formulations of an idea turn out to be basically-equivalent, that's strong evidence that we're on the right track.

In terms of applications, the locality results in particular are exactly the sort of thing which I've been looking for since my [last update on testing the Natural Abstraction Hypothesis](#). In combination with the [generalized Koopman-Pitman-Darmois theorem](#), they get us to the point where calculating abstractions from a base model is roughly-tractable, i.e. it can be done in something like $O(N^3)$ time with respect to the size of the base model. That still isn't quite efficient enough to handle *big* models, and unfortunately big models are exactly where we expect to see nontrivial results, so I'm still not *quite* at the point of running empirical tests. But it feels like the hard part is now past; there are some clear steps forward on the math, and I expect that those will basically close the gap to efficient calculation.

On the theoretical side, the first leg of the [Natural Abstraction Hypothesis](#) says:

For most physical systems, the information relevant “far away” can be represented by a summary much lower-dimensional than the system itself.

Assuming the proofs in this post basically hold up, and the loopholes aren't critical, I think this claim is now basically proven. There's still some operationalization to be done (e.g. the “dimension” of the summary hasn't actually been addressed yet), loopholes to close (e.g. deterministic computation makes things tricky), some legwork to flesh it all out (e.g. including numerical approximation), various extensions (e.g. logical uncertainty), and a lot of distillation needed, but I think this math is enough to conclude that the basic claim is probably true in worlds following local laws (like ours).

The basic idea is also useful even in nonlocal models, which I'll hopefully write about in the not-too-distant future. That form is more readily applicable to clustering-style applications, like e.g. recognizing “pencils” as a kind of object.

Appendices

Proof Sketch: Factorization

We'll use three facts. First, $P[X^\infty | M_{\text{resample}}, X^0]$ is invariant under resampling any variable - i.e. the distribution reaches an equilibrium as we run the resampling process. I won't actually prove that, because I don't expect that there's anything new involved; standard MCMC convergence proofs should largely carry over (allowing for conserved quantities, of course). Formally:

$$P[X^t | M_{\text{resample}}, X^0] = P[X^{t-1} | M_{\text{resample}}, X^0] \text{ as } t \rightarrow \infty$$

Second, the resampler is local:

$$P[X_i = x_i | M_{\text{resample}}, X^{t-1} = x] = P[X_i = x_i | M_{\text{base}}, X_{\neq i} = x_{\neq i}]$$

... and

$$P[X_i = x_i | M_{\text{base}}, X_{\neq i} = x_{\neq i}] = P[X_i = x_i | M_{\text{base}}, X_{nb(i)} = x_{nb(i)}]$$

... so

$$P[X_i^t = x_i | M_{\text{resample}}, X_{\neq i}^{t-1} = x_{\neq i}] = P[X_i^t = x_i | M_{\text{resample}}, X_{\text{nb}(i)}^t = x_{\text{nb}(i)}]$$

... where $X_{\text{nb}(i)}$ indicates the neighbors of i in the graph on which M_{base} factors.

Third, $X_{\neq i}^{t-1}$ screens off X^0 from X^t (thus the “Markov Chain” part of “Markov Chain Monte Carlo”).

$$\text{So, } P[X^t | M_{\text{resample}}, X_{\neq i}^{t-1}] = P[X^t | M_{\text{resample}}, X^0, X_{\neq i}^{t-1}].$$

Combining these three, we find

$$P[X_i^t | M_{\text{resample}}, X^0, X_{\neq i}^{t-1}] = P[X_i^t | M_{\text{resample}}, X^0, X_{\text{nb}(i)}^t] \text{ as } t \rightarrow \infty$$

... i.e. the neighbors which screen off X_i from everything else in the base model also screen off X_i^∞ from everything else in X^∞ , given X^0 . The [Hammersley-Clifford Theorem](#) tells us that this is a sufficient condition for $P[X^\infty | M_{\text{resample}}, X^0]$ to factor over the graph G (modulo taking some limits).

Note that the Hammersley-Clifford theorem applies to undirected graphical models. I won’t prove the extension of our theorem to Bayes Nets here because, again, there’s nothing particularly novel involved.

Proof Sketch: Resampler Telephone Theorem

Once we have locality of X^∞ given X^0 , this theorem is easy: it’s exactly like the Telephone Theorem, but on the distribution $P[X^\infty | M_{\text{resample}}, X^0]$ rather than $P[X | M_{\text{base}}]$. Because $P[X^\infty | M_{\text{resample}}, X^0]$ factors over the same graph as $P[X | M_{\text{base}}]$, we can pick any sequence of nonoverlapping nested Markov blankets $B_1 \dots B_n$ in the base model, carry them over directly to X^∞ , and apply the Telephone Theorem argument:

- The relationship between B_1 and B_{n+1} is mediated by B_n , so $\text{MI}(B_1, B_n | X^0) \geq \text{MI}(B_1, B_{n+1} | X^0)$. In other words, mutual information with B_1 decreases as we move outward through the layers.
- MI is nonnegative and decreasing, so it must approach a limit as $n \rightarrow \infty$.
- Once the MI is arbitrarily close to that limit, information about B_1 is arbitrarily perfectly conserved.
- Information is perfectly conserved only when the information is carried by a deterministic constraint, i.e. $P[B_1 | M_{\text{resample}}, X^0, B_n] = P[B_1 | M_{\text{resample}}, X^0, f_n(B_n)]$ where $f_n(B_n) = f_{n+1}(B_{n+1})$ with probability 1 for some f_n, f_{n+1} .

The key thing to notice here is the deterministic constraint $f_n(B_n) = f_{n+1}(B_{n+1})$. In the two-variable example, we saw that exactly this kind of information is conserved by our resampling process: when we resample variables in B_n , the constraint value is perfectly “remembered” by B_{n+1} , and when we resamples variables in B_{n+1} , the constraint value is perfectly “remembered” by B_n . Newly-generated sample values are forced to be perfectly consistent with the constraint value, so that information sticks around.

So, if $f_n(B_n) = f_{n+1}(B_{n+1})$ with probability 1, and the blankets B_n and B_{n+1} are nonoverlapping, then the value $f = f_n(B_n) = f_{n+1}(B_{n+1})$ is perfectly conserved when resampling. It's perfectly conserved by the variables in B_n when resampling a variable outside of B_n , and perfectly conserved by the variables in B_{n+1} when resampling a variable outside of B_{n+1} . So, conditional on information conserved by the resampling process (or, equivalently for purposes of the X^∞ distribution, conditional on X^0), the value of f is known; it does not give any information at all. So, conditional on quantities conserved by resampling, the mutual information $MI(B_1, B_n | X^0)$ must drop to zero in the limit of large n .

Ngo and Yudkowsky on scientific reasoning and pivotal acts

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

This is a transcript of a conversation between Richard Ngo and Eliezer Yudkowsky, facilitated by Nate Soares (and with some comments from Carl Shulman). This transcript continues the [Late 2021 MIRI Conversations](#) sequence, following [Ngo's view on alignment difficulty](#).

Color key:

Chat by Richard and Eliezer Other chat

14. October 4 conversation

14.1. Predictable updates, threshold functions, and the human cognitive range

[Ngo][15:05]

Two questions which I'd like to ask Eliezer:

1. How strongly does he think that the "shallow pattern-memorisation" abilities of GPT-3 are evidence for Paul's view over his view (if at all)
 2. How does he suggest we proceed, given that he thinks directly explaining his model of the chimp-human difference would be the wrong move?
-

[Yudkowsky][15:07]

1 - I'd say that it's some evidence for the Dario viewpoint which seems close to the Paul viewpoint. I think it's some evidence for the Dario viewpoint because Dario seems to be the person who made something like an advance prediction about it. It's not enough to make me believe that you can straightforwardly extend the GPT architecture to 3e14 parameters and train it on 1e13 samples and get human-equivalent performance.

[Ngo][15:09]

Did you make any advance predictions, around the 2008-2015 period, of what capabilities we'd have before AGI?

[Yudkowsky][15:10]

not especially that come to mind? on my model of the future this is not particularly something I am supposed to know unless there is a rare flash of predictability.

[Ngo][15:11]

1 - I'd say that it's some evidence for the Dario viewpoint which seems close to the Paul viewpoint. say it's some evidence for the Dario viewpoint because Dario seems to be the person who made something like an advance prediction about it. It's not enough to make me believe that you can straightforwardly extend the GPT architecture to 3e14 parameters and train it on 1e13 samples and get human-equivalent performance.

For the record I remember Paul being optimistic about language when I visited OpenAI in summer 20 But I don't know how advanced internal work on GPT-2 was by then.

[Yudkowsky][15:13]

2 - in lots of cases where I learned more specifics about X, and updated about Y, I had the experience looking back and realizing that knowing *anything* specific about X would have predictably produced a directional update about Y. like, knowing anything in particular about how the first AGI eats computation, would cause you to update far away from thinking that biological analogies to the computation consumed by humans were a good way to estimate how many computations an AGI needs to eat. you know lots of details about how humans consume watts of energy, and you know lots of details about how modern AI consumes watts, so it's very visible that these quantities are so incredibly different and go through so many different steps that they're basically unanchored from each other.

I have specific ideas about how you get AGI that isn't just scaling up Stack More Layers, which lead me to think that the way to estimate the computational cost of it is not "3e14 parameters trained at 1e1 ops per step for 1e13 steps, because that much computation and parameters seems analogous to human biology and 1e13 steps is given by past scaling laws", a la recent OpenPhil publication. But it seems to me that it should be possible to have the abstract insight that knowing more about general intelligence in AGIs or in humans would make the biological analogy look less plausible, because you wouldn't be matching up an unknown key to an unknown lock.

Unfortunately I worry that this depends on some life experience with actual discoveries to get something this abstract-sounding on a gut level, because people basically never seem to make abstract updates of this kind when I try to point to them as predictable directional updates?

But, in principle, I'd hope there would be aspects of this where I could figure out how to show that *any* knowledge of specifics would probably update you in a predictable direction, even if it doesn't seem best for Earth for me to win that argument by giving specifics conditional on those specifics actually being correct, and it doesn't seem especially sound to win that argument by giving specifics that are wrong.

[Ngo][15:17]

I'm confused by this argument. Before I thought much about the specifics of the chimpanzee-human transition, I found the argument "humans foomed (by biological standards) so AIs will too" fairly compelling. But after thinking more about the specifics, it seems to me that the human foom was in part caused by a factor (sharp cultural shift) that won't be present when we train AIs.

[Yudkowsky][15:17]

sure, and other factors will be present in AIs but not in humans

[Ngo][15:17]

This seems like a case where more specific knowledge updated me away from your position, contrary to what you're claiming.

[Yudkowsky][15:18]

eg, human brains don't scale and mesh, while it's far more plausible that with AI you could just run more of it

that's a huge factor leading one to expect AI to scale faster than human brains did

it's like communication between humans, but squared!

this is admittedly a specific argument and I'm not sure how it would abstract out to any specific argument

[Ngo][15:20]

Again, this is an argument that I believed less after looking into the details, because right now it's pre difficult to throw more compute at neural networks at runtime.

Which is not to say that it's a bad argument, the differences in compute-scalability between humans AIs are clearly important. But I'm confused about the structure of your argument that knowing more details will predictably update me in a certain direction.

[Yudkowsky][15:21]

I suppose the genericized version of my actual response to that would be, "architectures that have a harder time eating more compute are architectures which, for this very reason, are liable to need better versions invented of them, and this in particular seems like something that plausibly happens before scaling to general intelligence is practically possible"

[Soares][15:23]

(Eliezer, I see Richard as requesting that you either back down from, or clarify, your claim that any specific observations about how much compute AI systems require will update him in a predictable direction.)

[Ngo: 👍]

[Yudkowsky][15:24]

I'm not saying I know how to make that abstracted argument for exactly what Richard cares about, part because I don't understand Richard's exact model, just that it's one way to proceed past the point where the obvious dilemma crops up of, "If a theory about AGI capabilities is true, it is a disservice to Earth to speak it, and if a theory about AGI capabilities is false, an argument based on it is not sound"

[Ngo][15:25]

Ah, I see.

[Yudkowsky][15:26]

possible viewpoint to try: that systems in general often have threshold functions as well as smooth functions inside them.

only in ignorance, then, do we imagine that the whole thing is one smooth function.

the history of humanity has a threshold function of, like, communication or culture or whatever.

the correct response to this is not, "ah, so this was the unique, never-to-be-seen-again sort of fact which cropped up in the weirdly complicated story of humanity in particular, which will not appear in the much simpler story of AI"

this only sounds plausible because you don't know the story of AI so you think it will be a simple story

the correct generalization is "guess some weird thresholds will also pop up in whatever complicated story of AI will appear in the history books"

[Ngo][15:28]

Here's a quite general argument about why we shouldn't expect too many threshold functions in the impact of AI: because at any point, humans will be filling in the gaps of whatever AIs can't do. (The law of this type of smoothing is, I claim, why culture was a sharp threshold for humans - if there had been another intelligent species we could have learned culture from, then we would have developed more gradually.)

[Yudkowsky][15:30]

something like this indeed appears in my model of why I expect not much impact on GDP before AGI powerful enough to bypass human economies entirely

during the runup phase, pre-AGI won't be powerful to do "whole new things" that depend on doing lots of widely different things that humans can't do

just marginally new things that depend on doing one thing humans can't do, or can do but a bunch worse

[Ngo][15:31]

Okay, that's good to know.

Would this also be true in [a civilisation of village idiots?](#)

[Yudkowsky][15:32]

there will be sufficient economic reward for building out industries that are mostly human plus one that pre-AGI does, and people will pocket those economic rewards, go home, and not be more ambitious than that. I have trouble empathically grasping *why* almost all the CEOs are like this in our current Earth, because I am very much not like that myself, but observationally, the current Earth sure does seem to behave like rich people would almost uniformly rather not rock the boat too much.

I did not understand the whole thing about village idiots actually

do you want to copy and paste the document, or try rephrasing the argument?

[Ngo][15:35]

Rephrasing:

Claim 1: AIs will be better at doing scientific research (and other similar tasks) than village idiots, before we reach AGI.

Claim 2: Village idiots still have the core of general intelligence (which you claim chimpanzees don't have).

Claim 3: It would be surprising if narrow AI's research capabilities fell specifically into the narrow gap between village idiots and Einsteins, given that they're both general intelligences and are very similar in terms of architecture, algorithms, etc.

(If you deny claim 2, then we can substitute, say, someone at the 10th percentile of human intelligence - I don't know what specific connotations "village idiot" has to you.)

[Yudkowsky][15:37]

My models do not have an easy time of visualizing "as generally intelligent as a chimp, but specialized to science research, gives you superhuman scientific capability and the ability to make progress in most areas of science".

(this is a reference back to the pre-rephrase in the document)

it seems like, I dunno, "gradient descent can make you generically good at anything without that taking too much general intelligence" must be a core hypothesis there?

[Ngo][15:39]

I mean, we both agree that gradient descent can produce *some* capabilities without also producing much general intelligence. But claim 1 plus your earlier claims that narrow AIs won't surpass humans in scientific research, lead to the implication that the limitations of gradient-descent-without-much-general-intelligence fall in a weirdly narrow range.

[Yudkowsky][15:42]

I do credit the Village Idiot to Einstein Interval with being a little broader as a target than I used to think since the Alpha series of Go-players took a couple of years to go from pro to world-beating even once they had a scalable algorithm. Still seems to me that, over time, the wall clock time to traverse those ranges has been getting shorter, like Go taking less time than Chess. My intuitions still say that it'd be quite weird to end up hanging out for a long time with AGIs that conduct humanlike conversations and are ambitious enough to run their own corporations while those AGIs are still not much good at science.

But on my present model, I suspect the limitations of "gradient-descent-without-much-general-intelligence" to fall underneath the village idiot side?

[Ngo][15:43]

Oh, interesting.

That seems like a strong prediction

[Yudkowsky][15:43]

Your model, as I understand it, is saying, "But surely, GD-without-GI must suffice to produce better scientists than village idiots, by specializing chimps on science" and my current reply, though it's not a particular question I've thought a lot about before, is, "That... does not quite seem to me like a thing that should happen along the mainline?"

though, as always, in the limit of superintelligences doing things, or our having the Textbook From The Future, we could build almost any kind of mind on purpose if we knew how, etc.

[Ngo][15:44]

For example, I expect that if I prompt GPT-3 in the right way, it'll say some interesting and not-totally-nonsensical claims about advanced science.

Whereas it would be very hard to prompt a village idiot to do the same.

[Yudkowsky][15:44]

e.g., a superintelligence could load up chimps with lots of domain-specific knowledge they were not generally intelligent enough to learn themselves.

ehhhhhh, it is *not* clear to me that GPT-3 is better than a village idiot at advanced science, even in the narrow sense, especially if the village idiot is allowed some training

[Ngo][15:46]

It's not clear to me either. But it does seem plausible, and then it seems even more plausible that this will be true of GPT-4

[Yudkowsky][15:46]

I wonder if we're visualizing different village idiots

my choice of "village idiot" originally was probably not the best target for visualization, because in a
of cases, a village idiot - especially the stereotype of a village idiot - is, like, a damaged general
intelligence with particular gears missing?

[Ngo][15:47]

I'd be happy with "10th percentile intelligence"

[Yudkowsky][15:47]

whereas it seems like what you want is something more like "Homo erectus but it has language"

oh, wow, 10th percentile intelligence?

that's super high

GPT-3 is far far out of its league

[Ngo][15:49]

I think GPT-3 is far below this person's league in a lot of ways (including most common-sense reasoning)
but I become much less confident when we're talking about abstract scientific reasoning.

[Yudkowsky][15:51]

I think that if scientific reasoning were as easy as you seem to be imagining(?), the publication factor
of the modern world would be *much* more productive of real progress.

[Ngo][15:51]

Well, a 10th percentile human is very unlikely to contribute to real scientific progress either way

[Yudkowsky][15:53]

Like, on my current model of how the world really works, China pours vast investments into universities
and sober-looking people with PhDs and classes and tests and postdocs and journals and papers; but
none of this is the real way of Science which is actually, secretly, unbeknownst to China, passed down
rare lineages and apprenticeships from real scientist mentor to real scientist student, and China does
have much in the way of lineages so the extra money they throw at stuff doesn't turn into real science

[Ngo][15:52]

Can you think of any clear-cut things that they could do and GPT-3 can't?

[Yudkowsky][15:53]

Like... make sense... at all? Invent a handaxe when nobody had ever seen a handaxe before?

[Ngo][15:54]

You're claiming that 10th percentile humans invent handaxes?

[Yudkowsky][15:55]

The activity of rearranging scientific sentences into new plausible-sounding paragraphs is well within reach of publication factories, in fact, they often use considerably more semantic sophistication than that, and yet, this does not cumulate into real scientific progress even in quite large amounts.

I think GPT-3 is basically just Not Science Yet to a much greater extent than even these empty publication factories.

If 10th percentile humans don't invent handaxes, GPT-3 sure as hell doesn't.

[Ngo][15:55]

I don't think we're disagreeing. Publication factories are staffed with people who do better academically than 90+% of all humans.

If 90th-percentile humans are very bad at science, then of course GPT-3 and 10th-percentile humans are very very bad at science. But it still seems instructive to compare them (e.g. on tasks like "talk cogently about a complex abstract topic")

[Yudkowsky][15:58]

I mean, while it is usually weird for something to be barely within a species's capabilities while being within those capabilities at all, such that only relatively smarter individual organisms can do it, in the case of something that a social species has only very recently started to do collectively, it's plausible that the thing appeared at the point where it was barely accessible to the smartest members. Eg, it wouldn't be surprising if it would have taken a long time or forever for humanity to invent science from scratch, if all the Francis Bacons and Newtons and even average-intelligence people were eliminated leaving only the bottom 10%. Because our species just started doing that, at the point where our species was barely able to start doing that, meaning, at the point where some rare smart people could spearhead it, historically speaking. It's not obvious whether or not less smart people can do it over a longer time.

I'm not sure we disagree much about the human part of this model.

My guess is that our disagreement is more about GPT-3.

"Talk 'cogently' about a complex abstract topic" doesn't seem like much of anything significant to me. GPT-3 is 'cogent'. It fails to pass the threshold for inventing science and, I expect, for most particular sciences.

[Ngo][16:00]

How much training do you think a 10th-percentile human would need in a given subject matter (say, economics) before they could answer questions as well as GPT-3 can?

(Right now I think GPT-3 does better by default because it at least recognises the terminology, where most humans don't at all.)

[Yudkowsky][16:01]

I also expect that if you offer a 10th-percentile human lots of money, they can learn to talk more cogently than GPT-3 about narrower science areas. GPT-3 is legitimately more well-read at its lower level of intelligence, but train the 10-percentiler in a narrow area and they will become able to write better nonsense about that narrow area.

[Ngo][16:01]

This sounds like an experiment we can actually run.

[Yudkowsky][16:02]

Like, what we've got going on here is a real *breadth* advantage that GPT-3 has in some areas, but the breadth doesn't add up because it lacks the depth of a 10%er.

[Ngo][16:02]

If we asked them to read a single introductory textbook and then quiz both them and GPT-3 about it covered in that textbook, do you expect that the human would come out ahead?

[Yudkowsky][16:02]

AI has figured out how to do a subhumanly shallow kind of thinking, and it *is* to be expected that when AI can do anything at all, it can soon do more of that thing than the whole human species could do.

No, that's nothing remotely like giving the human the brief training the human needs to catch up to GPT-3's longer training.

A 10%er does not learn in an instant - they learn faster than GPT-3, but not in an instant.

This is more like a scenario of paying somebody to, like, sit around for a year with an editor, learning how to mix-and-match economics sentences until they can learn to sound more like they're making an argument than GPT-3 does, despite still not understanding any economics.

A lot of the learning would just go into producing sensible-sounding nonsense at all, since lots of 10%ers have not been to college and have not learned how to regurgitate rearranged nonsense for college teachers.

[Ngo][16:05]

What percentage of humans do you think could learn to beat GPT-3's question-answering by reading a single textbook over, say, a period of a month?

[Yudkowsky][16:06]

~_(ツ)_/~

[Ngo][16:06]

More like 0.5 or 5 or 50?

[Yudkowsky][16:06]

Humans cannot in general pass the Turing Test for posing as AIs!

What percentage of humans can pass as a calculator by reading an arithmetic textbook?

Zero!

[Ngo][16:07]

I'm not asking them to mimic GPT-3, I'm asking them to produce better answers.

[Yudkowsky][16:07]

Then it depends on what kind of answers!

I think a lot of 10%ers could learn to do wedding-cake multiplication, if sufficiently well-paid as adults rather than being tortured in school, out to 6 digits, thus handily beating the current GPT-3 at 'multiplication'.

[Ngo][16:08]

For example: give them an economics textbook to study for a month, then ask them what inflation is whether it goes up or down if the government prints more money, whether the price of something increases or decreases when the supply increases.

[Yudkowsky][16:09]

GPT-3 did not learn to produce its responses by reading *textbooks*.

You're not matching the human's data to GPT-3's data.

[Ngo][16:10]

I know, this is just the closest I can get in an experiment that seems remotely plausible to actually ru

[Yudkowsky][16:10]

You would want to collect, like, 1,000 Reddit arguments about inflation, and have the human read them and have the human produce their own Reddit arguments, and have somebody tell them whether the sounded like real Reddit arguments or not.

The textbook is just not the same thing at all.

I'm not sure we're at the core of the argument, though.

To me it seems like GPT-3 is allowed to be superhuman at producing remixed and regurgitated sentences about economics, because this is about as relevant to Science talent as a calculator being able to do perfect arithmetic, only less so.

[Ngo][16:15]

Suppose that the remixed and regurgitated sentences slowly get more and more coherent, until GPT-can debate with a professor of economics and sustain a reasonable position.

[Yudkowsky][16:15]

Are these points that GPT-N read elsewhere on the Internet, or are they new good points that no professor of economics on Earth has ever made before?

[Ngo][16:15]

I guess you don't expect this to happen, but I'm trying to think about what experiments we could run get evidence for or against it.

The latter seems both very hard to verify, and also like a very high bar - I'm not sure if most professo of economics have generated new good arguments that no other professor has ever made before.

So I guess the former.

[Yudkowsky][16:18]

Then I think that you can do this without being able to do science. It's a lot like if somebody with a really good memory was lucky enough to have read that exact argument on the Internet yesterday, & to have a little talent for paraphrasing. Not by coincidence, having this ability gives you - on my model no ability to do science, invent science, be the first to build handaxes, or design nanotechnology.

I admit, this does reflect my personal model of how Science works, presumably not shared by many leading bureaucrats, where in fact the papers full of regurgitated scientific-sounding sentences are not accomplishing much.

[Ngo][16:20]

So it seems like your model doesn't rule out narrow AIs producing well-reviewed scientific papers, since you don't trust the review system very much.

[Yudkowsky][16:23]

I'm trying to remember whether or not I've heard of that happening, like, 10 years ago.

My vague recollection is that things in the Sokal Hoax genre where the submissions succeeded, used humans to hand-generate the nonsense rather than any submissions in the genre having been purely machine-generated.

[Ngo][16:24]

Which doesn't seem like an unreasonable position, but it does make it harder to produce tests that we have opposing predictions on.

[Yudkowsky][16:24]

Obviously, that doesn't mean it couldn't have been done 10 years ago, because 10 years ago it's plausibly a lot easier to hand-generate passing nonsense than to write an AI program that does it.

oh, wait, I'm wrong!

<https://news.mit.edu/2015/how-three-mit-students-fooled-scientific-journals-0414>

In April of 2005 the team's submission, "Rooter: A Methodology for the Typical Unification of Access Points and Redundancy," was accepted as a non-reviewed paper to the World Multiconference on Systemics, Cybernetics and Informatics (WMSCI), a conference that Krohn says is known for "being spammy and having loose standards."

in 2013 IEEE and Springer Publishing removed more than 120 papers from their sites after a French researcher's analysis determined that they were generated via SCIGen

[Ngo][16:26]

Oh, interesting

Meta note: I'm not sure where to take the direction of the conversation at this point. Shall we take a brief break?

[Yudkowsky][16:27]

The creators continue to get regular emails from computer science students proudly linking to papers they've snuck into conferences, as well as notes from researchers urging them to make versions for other disciplines.

Sure! Resume 5p?

[Ngo][16:27]

Yepp

14.2. Domain-specific heuristics and nanotechnology

[Soares][16:41]

A few takes:

1. It looks to me like there's some crux in "how useful will the 'shallow' stuff get before dangerous things happen". I would be unsurprised if this spiraled back into the gradualness debate. I'm excited about attempts to get specific and narrow disagreements in this domain (not necessarily bettable; I nominal distilling out specific disagreements before worrying about finding bettable ones).
 2. It seems plausible to me we should have some much more concrete discussion about possible ways things could go right, according to Richard. I'd be up for playin the role of beeping when things seem insufficiently concrete.
 3. It seems to me like Richard learned a couple things about Eliezer's model in that last bout of conversation. I'd be interested to see him try to paraphrase his current understanding of it, and to see Eliezer produce beeps where it seems particularly off.
-

[Yudkowsky][17:00]



[Ngo][17:02]

Hmm, I'm not sure that I learned too much about Eliezer's model in this last round.

[Soares][17:03]

(dang :-p)

[Ngo][17:03]

It seems like Eliezer thinks that the returns of scientific investigation are very heavy-tailed.

Which does seem pretty plausible to me.

But I'm not sure how useful this claim is for thinking about the development of AI that can do science

I attempted in my document to describe some interventions that would help things go right.

And the levels of difficulty involved.

[Yudkowsky][17:07]

(My model is something like: there are some very shallow steps involved in doing science, lots of medium steps, occasional very deep steps, assembling the whole thing into Science requires having

the lego blocks available. As soon as you look at anything with details, it ends up 'heavy-tailed' because it has multiple pieces and says how things don't work if all the pieces aren't there.)

[Ngo][17:08]

Eliezer, do you have an estimate of how much slower science would proceed if everyone's IQs were shifted down by, say, 30 points?

[Yudkowsky][17:10]

It's not obvious to me that science proceeds significantly past its present point. I would not have the right to be surprised if Reality told me the correct answer was that a civilization like that just doesn't reach AGI, ever.

[Ngo][17:12]

Doesn't your model take a fairly big hit from predicting that humans just happen to be within 30 IQ points of not being able to get any more science?

It seems like a surprising coincidence.

Or is this dependent on the idea that doing science is much harder now than it used to be?

And so if we'd been dumber, we might have gotten stuck before newtonian mechanics, or else before relativity?

[Yudkowsky][17:13]

No, humanity is exactly the species that finds it barely possible to do science.

[Ngo][17:14]

It seems to me like humanity is exactly the species that finds it barely possible to do *civilisation*.

[Yudkowsky][17:14]

If it were possible to do it with less intelligence, we'd be having this conversation over the Internet than we'd developed with less intelligence.

[Ngo][17:15]

And it seems like many of the key inventions that enabled civilisation weren't anywhere near as intelligence-bottlenecked as modern science.

[Yudkowsky][17:15]

Yes, it does seem that there's quite a narrow band between "barely smart enough to develop agriculture" and "barely smart enough to develop computers"! Though there were genuinely fewer people in the preagricultural world, with worse nutrition and no Ashkenazic Jews, and there's the whole question about to what degree the reproduction of the shopkeeper class over several centuries was important to the Industrial Revolution getting started.

[Ngo][17:15]

(e.g. you'd get better spears or better plows or whatever just by tinkering, whereas you'd never get relativity just by tinkering)

[Yudkowsky][17:17]

I model you as taking a lesson from this which is something like... you can train up a villager to be Jol von Neumann by spending some evolutionary money on giving them science-specific brain features, since John von Neumann couldn't have been much more deeply or generally intelligent, and you could spend even more money and make a chimp a better scientist than John von Neumann.

My model is more like, yup, the capabilities you need to invent aqueducts sure do generalize the crap out of things, though also at the upper end of cognition there are compounding returns which can bring John von Neumann into existence, and also also there's various papers suggesting that selection was happening really fast over the last few millennia and real shifts in cognition shouldn't be ruled out. (last part is an update to what I was thinking when I wrote [Intelligence Explosion Microeconomics](#), and from my own perspective a more gradualist line of thinking, because it means there's a wider actual target to traverse before you get to von Neumann.)

[Ngo][17:20]

It's not that "von Neumann isn't much more deeply generally intelligent", it's more like "domain-specific heuristics and instincts get you a long way". E.g. soccer is a domain where spending evolutionary money on specific features will very much help you beat von Neumann, and so is art, and so is music

[Yudkowsky][17:20]

My skepticism here is that there's a version of, like, "invent nanotechnology" which routes through just the shallow places, which humanity stumbles over before we stumble over deep AGI.

[Ngo][17:21]

Would you be comfortable publicly discussing the actual cognitive steps which you think would be necessary for inventing nanotechnology?

[Yudkowsky][17:23]

It should not be overlooked that there's a very valid sibling of the old complaint "Anything you can do can be done by AI", which is that "Things you can do with surprisingly-to-your-model shallow cognition are precisely the things that Reality surprises you by telling you that AI can do earlier than you expected". When we see GPT-3, we were getting some amount of real evidence about AI capabilities advancing faster than I expected, and some amount of evidence about GPT-3's task being performable using shallower cognition than expected.

Many people were particularly surprised by Go because they thought that Go was going to require deeper real thought than chess.

And I think AlphaGo probably was thinking in a legitimately deeper way than Deep Blue. Just not as much deeper as Douglas Hofstadter thought it would take.

Conversely, people thought a few years ago that driving cars really seemed to be the sort of thing that machine learning would be good at, and were unpleasantly surprised by how the last 0.1% of driving conditions were resistant to shallow techniques.

Despite the inevitable fact that some surprises of this kind now exist, and that more such surprises will exist in the future, it continues to seem to me that science-and-engineering on the level of "invent nanotech" still seems pretty unlikely to be easy to do with shallow thought, by means that humanity discovers before AGI tech manages to learn deep thought?

What actual cognitive steps? Outside-the-box thinking, throwing away generalizations that govern your previous answers and even your previous questions, inventing new ways to represent your questions, figuring out which questions you need to ask and developing plans to answer them; these are some answers that I hope will be sufficiently useless to AI developers that it is safe to give them, while still pointing in the direction of things that have an un-GPT-3-like quality of depth about them.

Doing this across unfamiliar domains that couldn't be directly trained in by gradient descent because they were too expensive to simulate a billion examples of

If you have something this powerful, why is it not also noticing that the world contains humans? Why it not noticing itself?

[Ngo][17:30]

If humans were to invent this type of nanotech, what do you expect the end intellectual result to be?

E.g. consider the human knowledge involved in building cars

There are thousands of individual parts, each of which does a specific thing

[Yudkowsky][17:30]

Uhhhh... is there a reason why "Eric Drexler's *Nanosystems* but, like, the real thing, modulo however much Drexler did not successfully Predict the Future about how to do that, which was probably a lot" not the obvious answer here?

[Ngo][17:31]

And some deep principles governing engines, but not really very crucial ones to actually building (ea versions of) those engines

[Yudkowsky][17:31]

that's... not historically true at all?

getting a grip on quantities of heat and their flow was *critical* to getting steam engines to work

it didn't happen until the math was there

[Ngo][17:32]

Ah, interesting

[Yudkowsky][17:32]

maybe you can be a mechanic banging on an engine that somebody else designed, around principles that somebody even earlier invented, without a physics degree

but, like, engineers have actually needed math since, like, that's been a thing, it wasn't just a prestige trick

[Ngo][17:34]

Okay, so you expect there to be a bunch of conceptual work in finding equations which govern nanosystems.

Uhhhh... is there a reason why "Eric Drexler's *Nanosystems* but, like, the real thing, modulo however much Drexler did not successfully Predict the Future about how to do that, which was probably a lot" is not the obvious answer here?

This may in fact be the answer; I haven't read it though.

[Yudkowsky][17:34]

or other abstract concepts than equations, which have never existed before like, maybe not with a type signature unknown to humanity, but with specific instances unknown to present humanity
that's what I'd expect to see from humanly designed nanosystems

[Ngo][17:35]

So something like AlphaFold is only doing a very small proportion of the work here, since it's not able generate new abstract concepts (of the necessary level of power)

[Yudkowsky][17:35]

yeeeessss, that is why DeepMind did not take over the world last year
it's not just that AlphaFold lacks the concepts but that it lacks the machinery to invent those concept and the machinery to do anything with such concepts

[Ngo][17:38]

I think I find this fairly persuasive, but I also expect that people will come up with increasingly clever ways to leverage narrow systems so that they can do more and more work.
(including things like: if you don't have enough simulations, then train another narrow system to help that, etc)

[Yudkowsky][17:39]

(and they will accept their trivial billion-dollar-payouts and World GDP will continue largely undisturbed on my mainline model, because it will be easiest to find ways to make money by leveraging narrow systems on the less regulated, less real parts of the economy, instead of trying to build houses or do medicine, etc.)

real tests being expensive, simulation being impossibly expensive, and not having enough samples to train your civilization's current level of AI technology, is not a problem you can solve by training a new AI to generate samples, because you do not have enough samples to train your civilization's current level of AI technology to generate more samples

[Ngo][17:41]

Thinking about nanotech makes me more sympathetic to the argument that developing general intelligence will bring a sharp discontinuity. But it also makes me expect longer timelines to AGI, during which there's more time to do interesting things with narrow AI. So I guess it weighs more against Dario's view, less against Paul's view.

[Yudkowsky][17:41]

well, I've been debating Paul about that separately in the timelines channel, not sure about recapitulating it here

but in broad summary, since I expect the future to look like it was drawn from the "history book" barrel and not the "futurism" barrel, I expect huge barriers to doing *huge* things with narrow AI in small amounts of time; you can sell waifutech because it's unregulated and hard to regulate, but that does feed into core mining and steel production.

we could already have double the GDP if it was legal to build houses and hire people, etc., and the change brought by pre-AGI will perhaps be that our GDP could *quadruple* instead of just *double* if it were legal to do things, but that will not make it legal to do things, and why would anybody try to do things and probably fail when there are easier \$36 billion profits to be made in waifutech.

14.3. Relatively shallow cognition, Go, and math

[Ngo][17:45]

I'd be interested to see Paul's description of how we would train AIs to solve hard scientific problems. think there's some prediction that's like "we train it on arxiv and fine-tune it until it starts to output credible hypotheses about nanotech". And this seems like it has a step that's quite magical to me, b~~u~~ perhaps that'll be true of any prediction that I make before fully understanding how intelligence work

[Yudkowsky][17:46]

my belief is not so much that this training can never happen, but that this probably means the system was trained *beyond the point of safe shallowness*

not in principle over all possible systems a superintelligence could build, but in practice when it happens on Earth

my only qualm about this is that current techniques make it possible to buy shallowness in larger quantities than this Earth has ever seen before, and people are looking for surprising ways to make use of that

so I weigh in my mind the thought of Reality saying Gotcha! by handing me a headline I read tomorrow about how GPT-4 has started producing totally reasonable science papers that are actually correct

and I am pretty sure that exact thing doesn't happen

and I ask myself about GPT-5 in a few more years, which had the same architecture as GPT-3 but more layers and more training, doing the same thing

and it's still largely "nope"

then I ask myself about people in 5 years being able to use the shallow stuff *in any way whatsoever* to produce the science papers

and of course the answer there is, "okay, but is it doing that without having shallowly learned stuff that adds up to deep stuff which is *why it can now do science*"

and I try saying back "no, it was born of shallowness and it remains shallow and it's just doing science because it turns out that there is totally a way to be an incredibly mentally shallow skillful scientist if you think 10,000 shallow thoughts per minute instead of 1 deep thought per hour"

and my brain is like, "I cannot absolutely rule it out but it really seems like trying to call the next big surprise in 2014 and you guess self-driving cars instead of Go because how the heck would you guess that Go was shallower than self-driving cars"

like, that is an *imaginable* surprise

[Ngo][17:52]

On that *particular* point it seems like the very reasonable heuristic of "pick the most similar task" would say that go is like chess and therefore you can do it shallowly.

[Yudkowsky][17:52]

but there's a world of difference between saying that a surprise is imaginable, and that it wouldn't surprise you

[Ngo][17:52]

I wasn't thinking that much about AI at that point, so you're free to call that post-hoc.

[Yudkowsky][17:52]

the Chess techniques had already failed at Go

actual new techniques were required

the people around at the time had witnessed sudden progress on self-driving cars a few years earlier

[Ngo][17:53]

My advance prediction here is that "math is like go and therefore can be done shallowly".

[Yudkowsky][17:53]

self-driving cars were of obviously greater economic interest as well

my recollection is that talk of the time was about self-driving

heh! I have the same sense.

that is, math being shallower than science.

though perhaps not as shallow as Go, and you will note that Go has fallen and Math has not

[Ngo][17:54]

right

I also expect that we'll need new techniques for math (although not as different from the go techniques as the go techniques were from chess techniques)

But I guess we're not finding strong disagreements here either.

[Yudkowsky][17:57]

if Reality came back and was like "Wrong! Keeping up with the far reaches of human mathematics is harder than being able to develop your own nanotech," I would be like "What?" to about the same degree as being "What?" on "You can build nanotech just by thinking trillions of thoughts that are too shallow to notice humans!"

[Ngo][17:58]

Perhaps let's table this topic and move on to one of the others Nate suggested? I'll note that walking through the steps required to invent a science of nanotechnology does make your position feel more compelling, but I'm not sure how much of that is the general "intelligence is magic" intuition I mentioned before.

[Yudkowsky][17:59]

How do you suspect your beliefs would shift if you had any detailed model of intelligence?

Consider trying to imagine a particular wrong model of intelligence and seeing what it would say differently?

(not sure this is a useful exercise and we could indeed try to move on)

[Ngo][18:01]

I think there's one model of intelligence where scientific discovery is more actively effortful - as in, you need to be very goal-directed in determining hypotheses, testing hypotheses, and so on.

And there's another in which scientific discovery is more constrained by flashes of insight, and the systems which are producing those flashes of insight are doing pattern-matching in a way that's fairly disconnected from the real-world consequences of those insights.

[Yudkowsky][18:05]

The first model is true and the second one is false, if that helps. You can tell this by contemplating where you would update if you learned any model, by considering that things look more disconnected when you can't see the machinery behind them. If you don't know what moves the second hand on a watch and the minute hand on a watch, they could just be two things that move at different rates for completely unconnected reasons; if you can see inside the watch, you'll see that the battery is shared and the central timing mechanism is shared and then there's a few gears to make the hands move at different rates.

Like, in my ontology, the notion of "effortful" doesn't particularly parse as anything basic, because it doesn't translate over into paperclip maximizers, which are neither effortful nor effortless.

But in a human scientist you've got thoughts being shoved around by all sorts of processes behind the curtains, created by natural selection, some of them reflecting shards of Consequentialism / shadowy paths through time

The flashes of insight come to people who were looking in nonrandom places

If they didn't plan deliberately and looked on pure intuition, they looked with an intuition trained by past success and failure

Somebody walking doesn't plan to walk, but long ago as a baby they learned from falling over, and the ancestors who fell over more didn't reproduce

[Ngo][18:09]

I think the first model is probably more true for humans in the domain of science. But I'm uncertain about the extent to which this is true because humans have not been optimised very much for doing science - we consider the second model in a domain that humans have actually been optimised very hard for (say, physical activity) - then maybe we can use the analogy of a coach and a player. The coach can tell the player what to practice, but almost all the work is done by the player practicing in a way which updates their intuitions.

This has become very abstract, though.

14.4. Pivotal acts and historical precedents

[Ngo][18:11]

A few takes:

1. It looks to me like there's some crux in "how useful will the 'shallow' stuff get before dangerous things happen". I would be unsurprised if this spiraled back into the gradualness debate. I'm excited about attempts to get specific and narrow disagreements in this domain (not necessarily bettable; nominate distilling out specific disagreements before worrying about finding bettable ones).

2. It seems plausible to me we should have some much more concrete discussion about possible ways things could go right, according to Richard. I'd be up for playing the role of beeping when things seem insufficiently concrete.

3. It seems to me like Richard learned a couple things about Eliezer's model in that last bout of conversation. I'd be interested to see him try to paraphrase his current understanding of it, and to see Eliezer produce beeps where it seems particularly off.

Here's Nate's comment.

We could try his #2 suggestion: concrete ways that things could go right.

[Soares][18:12]

(I am present and am happy to wield the concreteness-hammer)

[Ngo][18:13]

I think I'm a little cautious about this line of discussion, because my model doesn't strongly constrain the ways that different groups respond to increasing developments in AI. The main thing I'm confident about is that there will be much clearer responses available to us once we have a better picture of AI development.

E.g. before modern ML, the option of international constraints on compute seemed much less salient, because algorithmic developments seemed much more important.

Whereas now, tracking/constraining compute use seems like one promising avenue for influencing AI development.

Or in the case of nukes, before knowing the specific details about how they were constructed, it would be hard to give a picture of how arms control goes well. But once you know more details about the process of uranium enrichment, you can construct much more efficacious plans.

[Yudkowsky][18:19]

Once we knew specific things about bioweapons, countries developed specific treaties for controlling them, which failed (according to @CarlShulman)

[Ngo][18:19, moved two down in log]

(As a side note, I think that if Eliezer had been around in the 1930s, and you described to him what actually happened with nukes over the next 80 years, he would have called that "insanely optimistic")

[Yudkowsky][18:21]

Mmmmmmaybe. Do note that I tend to be more optimistic than the average human about, say, global warming, or everything in transhumanism outside of AGI.

Nukes have going for them that, in fact, nobody has an incentive to start a global thermonuclear war. Eliezer is not in fact pessimistic about everything and views his AGI pessimism as generalizing to very few other things, which are not, in fact, as bad as AGI.

[Ngo][18:21]

I think I put this as the lowest application of competent power out of the things listed in my doc; I'd need to look at the historical details to know if important decision-makers actually cared about it, or were just doing it for PR reasons.

[Shulman][18:22]

Once we knew specific things about bioweapons, countries developed specific treaties for controlling them, which failed (according to @CarlShulman)

The treaties were pro forma without verification provisions because the powers didn't care much about bioweapons. They did have verification for nuclear and chemical weapons which did work.

[Yudkowsky][18:22]

But yeah, compared to pre-1946 history, nukes actually kind of did go *really surprisingly well!*

Like, this planet used to be a huge warring snakepit of Great Powers and Little Powers and then nuke came along and people actually got serious and decided to stop having the largest wars they could fi

[Shulman][18:22][18:23]

The analog would be an international agreement to sign a nice unenforced statement of AI safety principles and then all just building AGI in doomy ways without explicitly saying they're doing it..

The BWC also allowed 'defensive' research that is basically as bad as the offensive kind.

[Yudkowsky][18:23]

The analog would be an international agreement to sign a nice unenforced statement of AI safety principles and then all just building AGI in doomy ways without explicitly saying they're doing it..

This scenario sure sounds INCREDIBLY PLAUSIBLE, yes

[Ngo][18:22]

On that point: do either of you have strong opinions about the anthropic shadow argument about nukes? That seems like one reason why the straw 1930s-Eliezer I just cited would have been justified.

[Yudkowsky][18:23]

I mostly don't consider the anthropic shadow stuff

[Shulman][18:24]

In the late Cold War Gorbachev and Reagan might have done the BWC treaty+verifiable dismantling, but they were in a rush on other issues like nukes and collapse of the USSR.

Putin just wants to keep his bioweapons program, it looks like. Even denying the existence of the exposed USSR BW program.

[Yudkowsky][18:25]

I'm happy making no appeal to anthropics here.

[Shulman][18:25]

Boo anthropic shadow claims. Always dumb.

(Sorry I was only invoked for BW, holding my tongue now.)

[Yudkowsky: ♥] [Soares: ♥]

[Yudkowsky][18:26]

There may come a day when the strength of nonanthropic reasoning fails... but that is not this day!

[Ngo][18:27]

Okay, happy to rule that out for now too. So yeah, I picture 1930s-Eliezer pointing to technological trends and being like "by default, 30 years after the first nukes are built, you'll be able to build one in your back yard. And governments aren't competent enough to stop that happening."

And I don't think I could have come up with a compelling counterargument back then.

[Soares][18:27]

[Sorry I was only invoked for BW, holding my tongue now.]

(fwiw, I thought that when Richard asked "you two" re: anthropic shadow, he meant you also. But I appreciate the caution. And in case Richard meant me, I will note that I agree w/ Carl and Eliezer on this count.)

[Ngo][18:28]

(fwiw, I thought that when Richard asked "you two" re: anthropic shadow, he meant you also. But I appreciate the caution. And in case Richard meant me, I will note that I agree w/ Carl and Eliezer on this count.)

Oh yeah, sorry for the ambiguity, I meant Carl.

I do believe that AI control will be more difficult than nuclear control, because AI is so much more useful. But I also expect that there will be many more details about AI development that we don't currently understand, that will allow us to influence it (because AGI is a much more complicated concept than "really really big bomb").

[Yudkowsky][18:29]

[So yeah, I picture 1930s-Eliezer pointing to technological trends and being like "by default, 30 years after the first nukes are built, you'll be able to build one in your back yard. And governments aren't competent enough to stop that happening."]

And I don't think I could have come up with a compelling counterargument back then.]

So, I mean, in fact, I don't prophesize doom from very many trends at all! It's literally just AGI that is anywhere near that unmanageable! Many people in EA are more worried about biotech than I am, for example.

[Ngo][18:31]

I appreciate that my response is probably not very satisfactory to you here, so let me try to think about more concrete things we can disagree about.

[Yudkowsky][18:31]

[I do believe that AI control will be more difficult than nuclear control, because AI is so much more useful. But I also expect that there will be many more details about AI development that we don't

currently understand, that will allow us to influence it (because AGI is a much more complicated concept than "really really big bomb").]

Er... I think this is not a correct use of the Way I was attempting to gesture at; things being more complicated when known than unknown, does not mean you have more handles to influence them because each complication has the potential to be a handle. It is not in general true that very complicated things are easier for humanity in general, and governments in particular, to control, because they have so many exposed handles.

I think there's a valid argument about it maybe being more possible to control the supply chain for AI training processors if the global chip supply chain is narrow (also per Carl).

[Ngo][18:34]

One thing that we seemed to disagree on, to a significant extent, is the difficulty of "US and China preventing any other country from becoming a leader in AI"

[Yudkowsky][18:35]

It is in fact a big deal about nuclear tech that uranium can't be mined in every country, as I understand it, and that centrifuges stayed at the frontier of technology and were harder to build outside the well-developed countries, and that the world ended up revolving around a few Great Powers that had no interest in nuclear tech proliferating any further.

[Ngo][18:35]

It seems to me that the US and/or China could apply a lot of pressure to many countries.

[Yudkowsky][18:35]

Unfortunately, before you let that encourage you too much, I would also note it was an important fact about nuclear bombs that they did not produce streams of gold and then ignite the atmosphere if you turned up the stream of gold too high with the actual thresholds involved being unpredictable.

[Ngo][18:35]

E.g. if the UK had actually seriously tried to block Google's acquisition of DeepMind, and the US had actually seriously tried to convince them not to do so, then I expect that the UK would have folded. (Although it's a weird hypothetical.)

Unfortunately, before you let that encourage you too much, I would also note it was an important fact about nuclear bombs that they did not produce streams of gold and then ignite the atmosphere if you turned up the stream of gold too high with the actual thresholds involved being unpredictable.

Not a critical point, but nuclear power does actually seem like a "stream of gold" in many ways.

(also, quick meta note: I need to leave in 10 mins)

[Yudkowsky][18:38]

I would be a lot more cheerful about a few Great Powers controlling AGI if AGI produced wealth, but more powerful AGI produced no more wealth; if AGI was made entirely out of hardware, with no software component that could be keep getting orders of magnitude more efficient using hardware-independent ideas; and if the button on AGIs that destroyed the world was clearly labeled.

That does take AGI to somewhere in the realm of nukes.

[Ngo][18:38]

How much improvement do you think can be eeked out of existing amounts of hardware if people just to focus on algorithmic improvements?

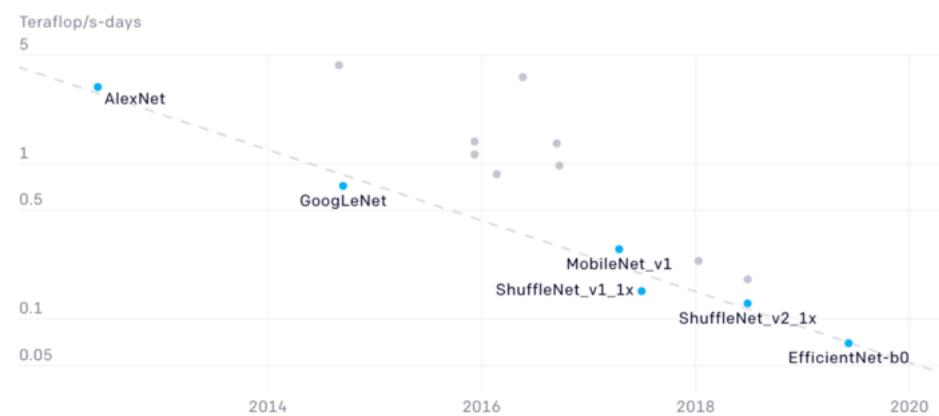
[Yudkowsky][18:38]

And Eliezer is capable of being less concerned about things when they are intrinsically less concerning which is why my history does not, unlike some others in this field, involve me running also being Terra Concerned about nuclear war, global warming, biotech, and killer drones.

[Ngo][18:39]

This says 44x improvements over 7 years: <https://openai.com/blog/ai-and-efficiency/>

44x less compute required to get to AlexNet performance 7 years later



[Yudkowsky][18:39]

Well, if you're a superintelligence, you can probably do human-equivalent human-speed general intelligence on a 286, though it might possibly have less fine motor control, or maybe not, I don't know

[Ngo][18:40]

(within reasonable amounts of human-researcher-time - say, a decade of holding hardware fixed)

[Yudkowsky][18:40]

I wouldn't be surprised if human ingenuity asymptoted out at AGI on a home computer from 1995. Don't know if it'd take more like a hundred years or a thousand years to get fairly close to that.

[Ngo][18:41]

Does this view cash out in a prediction about how the AI and Efficiency graph projects into the future

[Yudkowsky][18:42]

The question of how efficiently you can perform a fixed algorithm doing fixed things, often pales compared to the gains on switching to different algorithms doing different things.

Given government control of all the neural net training chips and no more public GPU farms, I buy that they could keep a nuke!AGI (one that wasn't tempting to crank up and had clearly labeled Doom-Causing Buttons whose thresholds were common knowledge) under lock of the Great Powers for 7 years during which software decreased hardware requirements by 44x. I am a bit worried about how long it takes before there's a proper paradigm shift on the level of deep learning getting started in 2006, after which the Great Powers need to lock down on individual GPUs.

[Ngo][18:46]

Hmm, okay.

14.5. Past ANN progress

[Ngo][18:46]

I don't expect another paradigm shift like that

(in part because I'm not sure the paradigm shift actually happened in the first place - it seems like neural networks were improving pretty continuously over many decades)

[Yudkowsky][18:47]

I've noticed that opinion around OpenPhil! It makes sense if you have short timelines and expect the world to end before there's another paradigm shift, but OpenPhil doesn't seem to expect that either.

Yeah, uh, there was kinda a paradigm shift in AI between say 2000 and now. There really, really was.

[Ngo][18:49]

What I mean is more like: it's not clear to me that an extrapolation of the trajectory of neural networks made much better by incorporating data about the other people who weren't using neural networks.

[Yudkowsky][18:49]

Would you believe that at one point Netflix ran a prize contest to produce better predictions of their users' movie ratings, with a \$1 million prize, and this was one of the largest prizes ever in AI and got tons of contemporary ML people interested, and neural nets were not prominent on the solutions list at all, because, back then, people occasionally solved AI problems *not using neural nets*?

I suppose that must seem like a fairy tale, as history always does, but I lived it!

[Ngo][18:50]

(I wasn't denying that neural networks were for a long time marginalised in AI)

I'd place much more credence on future revolutions occurring if neural networks had actually only been invented recently.

(I have to run in 2 minutes)

[Yudkowsky][18:51]

The world might otherwise end before the next paradigm shift, but if the world keeps on ticking for 1 years, 20 years, there will not always be the paradigm of training massive networks by even more massive amounts of gradient descent; I do not think that is actually the most efficient possible way to turn computation into intelligence.

Neural networks stayed stuck at only a few layers for a long time, because the gradients would explode or die out if you made the networks any deeper.

There was a critical moment in 2006(?) where Hinton and Salakhutdinov(?) proposed training Restricted Boltzmann machines unsupervised in layers, and then 'unrolling' the RBMs to initialize the weights in the network, and then you could do further gradient descent updates from there, because the activations and gradients wouldn't explode or die out given that initialization. That got people to, I dunno, 6 layers instead of 3 layers or something? But it focused attention on the problem of exploding gradients as the reason why deeply layered neural nets never worked, and that kicked off the entire modern field of deep learning, more or less.

[Ngo][18:56]

Okay, so are you claiming that that neural networks were mostly bottlenecked by algorithmic improvements, not compute availability, for a significant part of their history?

[Yudkowsky][18:56]

If anybody goes back and draws a graph claiming the whole thing was continuous if you measure the right metric, I am not really very impressed unless somebody at the time was using that particular graph and predicting anything like the right capabilities off of it.

[Ngo][18:56]

If so this seems like an interesting question to get someone with more knowledge of ML history than me to dig into; I might ask around.

[Yudkowsky][18:57]

[Okay, so are you claiming that that neural networks were mostly bottlenecked by algorithmic improvements, not compute availability, for a significant part of their history?]

Er... yeah? There was a long time when, even if you threw a big neural network at something, it just wouldn't work.

Good night, btw?

[Ngo][18:57]

Let's call it here; thanks for the discussion.

[Soares][18:57]

Thanks, both!

[Ngo][18:57]

I'll be interested to look into that claim, it doesn't fit with the impressions I have of earlier bottlenecks. I think the next important step is probably for me to come up with some concrete governance plans that I'm excited about.

I expect this to take quite a long time

[Soares][18:58]

We can coordinate around that later. Sorry for keeping you so late already, Richard.

[Ngo][18:59]

No worries

My proposal would be that we should start on whatever work is necessary to convert the debate into publicly accessible document now

In some sense coming up with concrete governance plans is my full-time job, but I feel like I'm still quite a way behind in my thinking on this, compared with people who have been thinking about governance specifically for longer

[Soares][19:01]

(@RobBensinger is already on it 😊)

[Bensinger:]

[Yudkowsky][19:03]

Nuclear plants might be like narrow AI in this analogy; some designs potentially contribute to proliferation, and you can get more economic wealth by building more of them, but they have no Unlabeled Doom Dial where you can get more and more wealth out of them by cranking them up until some unlabeled point the atmosphere ignites.

Also a thought: I don't think you just want somebody with more knowledge of AI history, I think you might need to ask an actual old fogey *who was there at the time*, and hasn't just learned an ordered history of just the parts of the past that are relevant to the historian's theory about how the present happened.

Two of them, independently, to see if the answers you get are reliable-as-in-statistical-reliability.

[Soares][19:19]

My own quick take, for the record, is that it looks to me like there are two big cruxes here.

One is about whether "deep generality" is a good concept, and in particular whether it pushes AI systems quickly from "nonscary" to "scary" and whether we should expect human-built AI systems to acquire it in practice (before the acute risk period is ended by systems that lack it). The other is about how easy it will be to end the acute risk period (eg by use of politics or nonscary AI systems alone).

I suspect the latter is the one that blocks on Richard thinking about governance strategies. I'd be interested in attempting further progress on the former point, though it's plausible to me that that should happen over in #timelines instead of here.

My attitude towards death

The philosophy and psychology of death seem weirdly under-discussed - particularly by the wider silicon valley community, given how strongly anti-death many people in it are. This post is an attempt to think through some relevant considerations, primarily focused on my own intuitions and emotions. See also this [old blog post](#) - I mostly still agree with the points I made in it, but when thinking about it now I frame things pretty differently.

Fearing death, loving life

Let's first distinguish two broad types of reasons for wanting to avoid death: fearing death, and loving life.^[1] Perhaps these seem like two sides of the same coin - but, psychologically speaking, they feel very distinct to me. The former was particularly dominant when I was in primary school, when a part of me emerged that was very afraid of death (in a way that wasn't closely linked to fear of missing out on any particular aspects of life). That part is still with me - but when it comes to the surface, its fear feels viscerally unpleasant, so I learned to suppress it pretty strongly.

Arguments for why death is bad usually focus on positive reasons - living longer allows people to experience more happiness, and more of the other good things in life. These have resonated with me more over time, as I started to think about death on a more intellectual level. However, one difficulty with these arguments is that many parts of me pursue goals in a fairly myopic way which doesn't easily extrapolate to centuries, millennia, or longer. For example, it's hard to imagine what career success or social success look like on the scale of millennia - and even when I try, those visions are pretty different from the versions of those concepts that I currently find motivating on a gut level. Extrapolating hedonistic goals is easier in some ways (it's easy to imagine being happy for a very long time) but harder in other ways (the parts of me which care most about happiness are also the most myopic).

Dissolving fear

In practice, then, most of my motivation for avoiding death in the long term stems from fear of death. Although that fear comes out only rarely, I have a strong heuristic that fear-based motivation should be transmuted to other forms of motivation wherever possible. So what would happen if I talked more to the part that's scared of death, to try and figure out where it's coming from? By default, I expect it'd be uncooperative - it wants to continue being scared of death, to make sure that I act appropriately (e.g. that I stay ambitious). Can I assure it that I'll still try hard to avoid death if it becomes less scared? One source of assurance is if I'm very excited about a very long life - which I am, because [the future could be amazing](#). Another comes from the altruistic part of me, whose primary focus is increasing the probability that the future will in fact be amazing. Since I believe that we face significant existential risk this century,^[2] working to make humanity's future go well overlaps heavily with working to make my own future go well. I think this broad argument (along with being in communities which reward longtermist altruism) has helped make the part of me that's scared of death more quiescent.

Indeed, probably my main concern with my current attitude towards death is actually that I'm not scared enough about existential risk - I think that, if my emotions better

matched my credences, that'd help motivate me (especially to pursue particularly unusual or ambitious interventions). This doesn't seem like a crucial priority, though, since my excitement- and interest-based motivations have been working fairly well so far (modulo some other productivity gaps which seem pretty orthogonal).

Generalising to others

So far I've talked primarily about my own experience. I'm curious about how well this generalises to other people. It seems like fear of death is a near-universal emotion (it's striking that the [first recorded story we have](#) is about striving to escape death), but my guess is that most people have it much less strongly than I did.

Since most people aren't very openly concerned with avoiding death in the long term, I feel uncertain about the extent to which they've suppressed versus dissolved that fear. My guess is that in western societies most people have mainly suppressed it, and that the hostility they often show to longevity research or cryonics is a psychological defense mechanism. If so, then overcoming those defense mechanisms to convince people that death is not inevitable might unlock a lot of suppressed excitement about the future. However, I'm wary of assuming that other people are too similar to me - perhaps other people's fear of death is just more myopic than mine.

There also seem to be some people who started off with a long-term fear of death, then dissolved it, usually by significantly changing their conception of personal identity - via meditation, or drugs, or philosophical argument. The big question is whether this change is more like an empirical update, or more like a value shift (to be clear, I don't think that there's a bright line between the two - but something can be much more like one or the other). If the former, then perhaps fear of death is just a "mistake" that many people make. Whereas the latter suggests that death is really bad according to some people's values, and mostly fine for others, even though they may in other ways be psychologically similar. Both of these conclusions seem a bit weird; let's try to get a bit more clarity by digging into arguments about personal identity now.

Continuity of self

The core question is how much we should buy into the folk view of personal identity - the view that there's a single "thread" of experience which constitutes my self, where I survive if that thread continues and "die" if it breaks. I consider thought experiments about duplicates to provide strong evidence against this position - it seems very compelling to me that, when two identical copies of myself are created, there is no fact of the matter about which one is "really me". Insofar as many people have intuitions weighing the other way, that's probably because we evolved in an environment where identical duplication didn't happen. In a future where duplication exists, and we continue being subject to evolution, I can easily imagine the mental concept of survival-of-self being straightforwardly replaced by the concept of survival-of-a-copy-of-myself.

The main alternative to caring about continuity is caring about level of similarity - identifying with a successor if they are sufficiently psychologically similar to you. This might leave you identifying with many successors, or ones that are very disconnected from you in time or space. However, it's also consistent with identifying only with successors with a level of similarity that, in practice, will only be achievable by

copying or uploading you^[3] (although I expect that really buying into the similarity theory of personal identity will make most people more altruistic, [like it did for Parfit](#)).

The strongest argument in favour of the folk view arises when considering large universes, like quantum multiverses or spatially infinite universes. In a quantum multiverse there are many copies of myself, and I tend to experience being the ones with more measure. But what does that even mean? If I expect that N slightly different copies of myself will branch off soon, and all of them will have the experience of being me, how can I anticipate being more likely to “find myself” as a given one of them? There’s something here which I don’t understand, and which makes me hesitant to fully dismiss the idea of a thread of experiences (a confusion which Yudkowsky explores in [these two](#) posts). I think the appropriate response is to be cautious until we understand this better - for instance, I would currently strongly prefer being non-destructively rather than destructively uploaded.

Generalising to society

When we stop thinking on an individual level and start thinking on a societal level, many more pragmatic considerations arise - especially related to how widespread longevity might shift the overall balance of power in the world. I do think these are important; here, though, I want to focus on a couple of broader philosophical considerations.

I previously talked about the part of myself which wants to make the future amazing. Partly that stems from [imagining all the different ways](#) in which the world might dramatically improve, including defeating death. Partly it’s an aesthetic preference about the trajectory of humanity - I want us to flourish in an analogous way to how I want to live a flourishing life myself. But there’s also a significant utilitarian motivation - which is relevant here because utilitarianism doesn’t care about death for its own sake, as long as the dead are replaced by new people with equal welfare. Indeed, if our lives have diminishing marginal value over time (which seems hard to dispute if you’re taking our own preferences into account at all), and humanity can only support a fixed population size, utilitarianism actively prefers that older people die and are replaced.

Now, I don’t think we’ll hit a “fixed population size” constraint until well after we’re posthuman, so this is a pretty abstract consideration. By that point, hopefully we won’t need to bite any bullets - we could build a flourishing civilisation which extrapolates our more human-specific values as well as possible, and also separately build the best utilitarian civilisation (assuming we can ensure non-conflict between them). But I’m also open to the idea that the future will look sufficiently weird that many of the concepts I’ve been using break down. For example, the boundaries between different minds could blur to such an extent that talking about the deaths of individuals doesn’t make much sense any more. I find it hard to viscerally desire that for myself, and I expect that most people alive today are much less open to the possibility than I am, but I can imagine changing my mind as we come to understand much more about how minds and values work.

1. ^

Upon reflection, I might also add a third distinct motivation - the celebration of immortality. I get this feeling particularly when I read fiction with very long-lived

characters. But since it's much weaker than the other two, I won't discuss it further.

2. ^

At least double digit percentage points, although my specific estimate is pretty unstable.

3. ^

On a side note: I feel very uncertain about how much information about my brain (in the form of my blog posts, tweets, background information about my life, etc) would be sufficient for future superintelligences to recreate me in a way that I'd consider a copy of myself. I haven't even seen any rough bounds on this - maybe worth looking into.

Observation

Knowing the territory takes patient and direct observation.

Imagine that you meet someone you're attracted to at a party. At one point, they smile at you, and you notice. You're pretty sure they like you, but you really want to know whether they *like* you like you.

You don't act on this in any particular way, but you do spend the whole next week thinking about it. You think about other people who have been into you, and about people who have not, and the differences between them. You muse about what sort of taste in romantic partners you imagine the person might have. By the end of the week, you're weighing your virtues and vices, trying to decide whether you're even worthy of love.

(If this seems alien to you, I hope it is at least true to your experiences of *some humans*.)

In the moment when you noticed you were attracted to the person, you made an observation. In the moment when you noticed their smile, you made another. In the moment when you noticed your curiosity, you made another.

But as soon as you vanished into your own musings, you were no longer making observations. You were no longer *collecting data*. Instead, you were interpolating, extrapolating, filling in the gaps with stories and guesses, processing and reprocessing. Everything that followed, in the week after the party, took place inside your map—analysis, interpretation, reasoning, reflection.

In Arthur Conan Doyle's "A Scandal in Bohemia," Sherlock Holmes lectures Watson on the difference between *seeing* and *observing*:

"You see, but you do not observe. The distinction is clear. For example, you have frequently seen the steps which lead up from the hall to this room."

"Frequently."

"How often?"

"Well, some hundreds of times."

"Then how many are there?"

"How many? I don't know."

"Quite so! You have not observed. And yet you have seen. That is just my point. Now, I know that there are seventeen steps, because I have both seen and observed."

I don't know how many steps there are on the staircase up to my own living room, either. Setting aside the question of prioritization, and whether I *should* be turning my attention there—what is it, exactly, that Watson and I are doing with the steps?

My guess is that we've taken some initial impressions—a few moments of impact from the external world—and used those points to draw a constellation. Every time we walk up the steps, we do almost all of our processing *on the constellation*, rather than on the points of light in the sky.

Most of our “seeing” the stairs is happening inside of our maps. We observe just enough to recognize that we’re about to encounter the well-understood “stairs” entity, and then we superimpose our “stairs” concept over whatever sensations are happening to us, and *stop paying attention*. To the extent that our brains record anything, it’s that we “climbed up the stairs,” rather than that we felt some number of impacts under each of our feet, while the muscles in our legs contracted and our heart rate climbed slightly, etc.



Imagine that you *do* end up asking the cute person from the party to meet you for coffee, but when the day comes, you're extremely distracted by a disaster at work, one you'll have to return to as soon as the date is over. Despite a whole hour of conversation, you leave feeling like you've learned almost nothing about them.

Crucial data was all around you, but while you *saw* it, you failed to *observe* any of it.

It is hardest to make fresh observations about things you have seen many times. The stairs, long-held beliefs, attitudes you were raised with. The more often you superimpose your drawing of a constellation over points of light in the sky, the more opaque your drawing becomes.

It probably doesn't really matter that I have seen-but-failed-to-observe my stairs. I never miss a step, and I'm not in a murder mystery whose solution might depend on how many steps there are.

It certainly *does* matter, though, if I have seen-but-failed-to-observe the way I make requests of my child, especially if I haven't even noticed the distinction. If I believe I've observed when I've really only seen, I'm much less likely to start paying attention, or to hypothesize that I may have gotten something wrong. If we're going to be close for a long time, we need to be able to communicate with *each other*, not just with the cartoon drawings we habitually plaster over each other's faces.

It also matters if I've seen-but-failed-to-observe the factors that cause me to continue on my current career path, what I count as evidence, or my default response when my expectations are violated.

Seeing-but-not-observing is a failure to make contact with a bit of territory that is right in front of you. It is standing at the bank of a river while staring at the part of your map labeled "river". Often that's good enough; but sometimes the river is flooded when you need to cross, and then you really have to lower your map and *make contact* with crucial data. You have to look at the world itself, or else you'll drown.

In the sentence "Knowing the territory takes patient and direct observation," this is what I mean by "observation." I mean *actual contact with the territory*. Looking at the stars themselves, instead of letting the constellation fill your mind as your eyes glaze over.

Knowing the territory takes patient and direct *contact with the territory*.

In the next two essays, I'll talk about two ways of being in contact with the territory: directly, and patiently.

Before Colour TV, People Dreamed in Black and White

On an episode of [Julia Galef's podcast](#), the philosopher Eric Schwitzgebel said the following:

For [dreams], there was actually a literature that's very interesting where people in the '50s in the United States and the '40s thought that dreams just generally were black and white. I don't think that they thought it was just dreams in the United States, as influenced by media. I think they just thought dreams are a black and white kind of thing. Most people thought that in the 1950s. It's related to the presence of media in the culture, so if you look pre-20th century, very few people will say that dreams are black and white. If you look 21st century, very few people will say that dreams are black and white. You look at the arc of it and it relates to the dominance of black and white film media in the culture.

And we got some cross-cultural evidence for this. This guy emailed me and said, "We should try this in China," because this was about the year 2000. He said, "Well, in rural China, most people are exposed to black and white media, their TVs are black and white, whereas in urban China, most people -- especially the wealthier people -- are exposed to mostly colour media." So we asked about their dreams and we found rural people in China in the early 2000s tended to say that their dreams were black and white, and urban people tended to say their dreams were coloured.

That became the paper [Schwitzgebel, Huang and Zhou 2006](#). If true, this is one of the most bonkers things I have ever learned.

The thing is, it's extremely unlikely that black and white TV actually changed the contents of people's dreams. There's no plausible way that the small proportion of time people spent watching visual media could radically change dreams about things we see in colour every day. Rather, *people don't know whether they dream in colour*. Dreams may not even have associated colours one way or the other! Indeed, when I asked a few friends and family whether they dreamed in colour, a surprising number of them answered "I don't know". When the dominant culture has a reference of visual media in black and white, you think you dream in black and white. And when your culture has a reference of visual media in colour, you think you dream in colour.

This relates to a generally underappreciated aspect of consciousness: *vagueness*. Your conscious experience of the world is vague. You don't typically know what you're feeling, or dreaming, and look to cultural cues to figure it out. This explains the stylised fact that anxiety and excitement are almost neurologically indistinguishable; the difference is in the surrounding interpretation. More speculatively, it also may explain the [cross-cultural differences in mental illnesses](#). The associated brain states of mental illnesses may well be the same everywhere, caused by a few failure modes. But different cultures prime people to think of mental illnesses in different ways.

You may be sceptical if you are aware of how the psychological research on priming [has not replicated well](#). But my colloquial usage of the term 'prime' is different from its technical meaning in psychology. It is not quite the placebo effect either: since all experiences are influenced by beliefs and expectations, that would commit us to say

that everything is a placebo, which doesn't seem right. It's more similar to the Popperian case against empiricism that I outlined in my [review of The Beginning of Infinity](#).

I was thinking out loud with a friend recently about how the purpose of meditation may be to eliminate this mental vagueness. To better understand sensation, unmediated by concepts. I heard [Sam Harris](#) say that experienced meditators even practice mindfulness in their sleep. It would be interesting to gather together people who claim to be enlightened and see if *they* dream in colour. Then again, monks probably don't watch a lot of TV.

Reward Good Bets That Had Bad Outcomes

Introduction

I am a very anxious person. One of the most damaging ways this manifests is that I am pretty risk-averse and afraid of failure. In situations of uncertainty, I often want to freeze up, and know I'll feel safer doing nothing.

This is a really bad problem! If I let this dictate my life, I lose a *ton* of value. In particular, there are a lot of areas in my life that are hits-based, where the best way to be successful is to persevere through many failures and [seek the upside risk](#) of things occasionally going *super* well. I want to be someone who can be a great researcher, find really awesome friends, and [generally be ambitious](#) about all areas of my life going well. And to achieve this, it is important that I be the kind of person who can take actions with high expected value, and persevere through failures without feeling paralysed. The key thing going wrong here is that I beat myself up over bad *outcomes* even when I had no way of knowing it wouldn't work out, given the information at the time. And anxiety gives me a negative prior and causes suboptimal outcomes to stick in my mind. Failures feel painful in a way that missed opportunities do not.

My solution to this is to **think in bets, not outcomes**. To clearly notice all of the good *bets* that I take, the actions that I endorse given what I knew at the time, and to be happy about each of those. And to think of my life in terms of this, rather than the concrete outcomes of the bet, and whether *that* was a success or failure. At the end of the day, the only thing I can control is the bets that I make, and the policies I follow, and there will always be uncertainty on the outcomes. And if I make a good bet with a bad outcome, I should be happy about this, not sad! I refuse to let my negative emotions be tied to things fundamentally outside my control.

My Underlying Model

I first formed this view when I did a trading internship a few years ago. In settings like financial trading or poker, the fundamental skill is about engaging well with uncertainty, and getting past anxieties is a key part of that! And noticing all of the expected value I was missing when I froze up did a lot for helping me notice this failure mode and learning how to solve it. And though the lessons generalise, real life is often a much harder learning environment than these settings - I make fewer bets and so get fewer data points, and it's much harder to explicitly calculate what's going on.

I find it easiest to understand what's going on here and how to fix it when thinking of myself as having a reinforcement learning system inside my head, shaping my actions. I take actions in the world, get feedback from my environment, and use this to update the policies that I follow. Within this framing, there are two clear problems with learning to make good bets with high upside, while being anxious.

The first problem arises because reality is noisy! Even if I had zero anxiety, fundamentally reality has unknowns and I must make decisions under uncertainty.

But, by default, I only learn about my actions from their outcomes. And this makes it *really* hard to learn strategies around pursuing occasional major upsides! It's obviously worth it to go on 99 unsuccessful dates if the hundredth results in marriage. But by default, my reinforcement learner will likely be discouraged and stop after 99 failures. While, if I can reframe it as 99 successful *bets*, then I get much better!

The second problem comes from anxiety, which causes me to *over-update* on negative feedback, and consider it *way* more important than positives. This is a fundamental issue with my learning algorithm that means I will learn systematically bad policies. By focusing on the action I took being good, this reduces the anxiety caused by unsuccessful outcomes. Note that negative feedback doesn't just include stuff that may actually be a big deal, like a romantic rejection, or missing out on a job I really cared about. At least for me, my anxiety reacts badly to even minor negative outcomes with no real consequences, like making a joke that didn't land, or recommending a book to someone that they've already read.

By default, I feel like I can solve these issues if I just [try harder](#). Think harder about an issue, go through every consideration, analyse it more deeply, and only take the actions that will work out well. This is an illusion! Reality is not fully knowable. And thinking harder has costs. If I follow the strategy of "just try harder", I will implicitly miss a lot of bets worth taking. The optimal strategy, given that I am an imperfect person in an uncertain world, is to take positive expected value bets. Finding ways to learn well in spite of anxiety is essential, because anxiety holds me back from so many bets worth taking.

How to Apply This?

The idea of thinking of my actions as bets and not focusing on their outcomes is a pretty core part of how I think about my life, and is useful in a wide range of areas in different ways. A quick brainstorm of different areas where my anxiety significantly holds me back from making the right bets:

- Applying for jobs
- Asking people out/going on dates
- Pursuing research directions
- Making friends, and generally [taking social initiative](#)
- Offering people help and favours
- Recommending books/articles/resources
- Introducing people who might get on
- Writing a blog post
 - Or starting a blog in the first place!
- Writing a cold email
- Giving advice
- [Asking for help](#)
 - Especially asking someone for their time, or anything else with a risk of rejection and that might be being a burden!
- Any form of [seeking upside risk](#)
- Sharing opportunities - I personally try hard to message people with jobs that might be a good fit, good articles I read they might enjoy, etc.
 - The instance of this I'm most proud of is getting stressed about Omicron near the start of the surge, and messaging 100 friends with instructions on how to get boosters earlier - this felt stressful at the time, but resulted in

5-10 counterfactually getting it a week or two earlier, and 1-3 getting a booster at all.

Exercise: Set a 5 minute timer and brainstorm times in *your* life when this bias applies. How could you orient to these in terms of bets, not outcomes?

The exact way I try to think in bets not outcomes varies depending on context, but there are a few core principles that stand out:

- Find ways to actively be excited about unsuccessful outcomes, so long as I think it was a good bet!
 - One way that works well for me is to reflect on how the action fits my self-identity, and is an example of becoming the kind of person I want to be. This successfully shifts focus from outcomes because my identity is a function of the actions I take, not the feedback from the world
 - This is the core insight of [becoming a person who actually does things](#)
 - Another way is to quantify things
 - Make a log of your unsuccessful bets, eg a list of rejections or failures. Set targets for how many failures you want to have, and see each one as an example of becoming someone who can put yourself out there!
 - Estimate the probability of the outcome you want! Eg [Chris Olah's framing of dating and meeting potential partners in terms of micro-marriages](#)
- Magnify your excitement about positive outcomes! Remember them, cherish them, and use them as motivation!
 - Keep a log of great outcomes, and bets going well.
 - Eg, I often share opportunities in group chats, and know of at least two people who've gotten internships this way - I find this super motivating to do it more often!
 - Relatedly, I keep a log of particularly happy memories and meaningful compliments, which is really uplifting to read when I'm down
 - Notice selection bias - all the good outcomes you might not hear about!
 - Eg, I occasionally hear from people who've had significant life improvements from things I've done. This is fucking awesome in and of itself, but even better when I reflect on how I likely miss out on most things like this!
 - Try to shift the selection bias, by making it clear that you love hearing about things like this, and being easy to reach!
 - If anything I've done has improved your life, [I'd love to hear about it!](#)
- Reflect on whether I *could* have done something differently, given what I knew at the time. This really helps to defuse the anxiety that I'm missing an important lesson and could have known better, and occasionally get the insight that it *was* a bad outcome!
 - It's important to focus on *given what I knew at the time*. If I'm not careful, my anxieties *love* to smuggle in some hindsight bias, and tell me that I'm an idiot for not having known the future! By focusing on general policies I could follow, I can get past this.
 - [Engage my inner simulator](#) and ask myself "Suppose, at the time, I predicted it would go badly and decided not to do it. Am I surprised by this outcome?"

- Further, ask myself what happened, and *why* I decided not to do it.
Was it for the right reasons?
- Take the outside view - is there any similar past action that did go well?
And if so, was this case obviously worse than that one, given what I knew beforehand back then? Can I find a policy which avoided this failure without missing out on that success?

Does This Reward Bad Bets Too?

One caveat worth addressing is whether this strategy could be *dangerous*? When I think about putting it into practice, this is the biggest flinch from my anxiety - maybe my bets are actually systematically bad and I am deluding myself, and the outcomes are the only way to get this feedback. This is obviously worth considering, and will sometimes happen! The ideal world is one where I evaluate each outcome for information that I'm missing, and take it as a slight negative update on whether the bet was worth making.

But there is no way of reaching that ideal world - my anxiety is a major bias, it pushes me towards risk-aversion, and it's basically impossible to *perfectly* correct for a bias like this. My solution essentially introduces [a counter-bias](#), towards ignoring the outcomes by default, which pushes me towards risk-seeking. In principle, there's some risk of overshooting the ideal point and being *too* risk-seeking, but in practice I think this is really unlikely! Especially if I explicitly reflect on whether I *could* have known better. My anxiety creates a pretty big bias towards risk-aversion, and dealing with anxiety is hard, and nothing I do is likely to create as big a bias the other way. I'm not able to ignore the anxiety at particularly bad outcomes, or the creeping doubt of getting way more unsuccessful outcomes than expected. The sheer fact that I feel anxiety about overshooting is a sign that I am safe, and can trust myself to not go too far without needing to actively track it!

Rewarding Other People's Bets

Many of my anxieties are social in nature, and I get way more anxious about bad outcomes involving other people. And it's much easier for someone *else* to help me overcome a socially-related anxiety by giving reliable feedback, than trying to deal with it within the insecurities of my own head. I like to seek positive externalities, so a great (and sad) thing is that this works in reverse - social anxieties are super common, so if I can help *other* people reward themselves for good bets with bad outcomes, I can help them make much better bets!

Often I do this by being enthusiastic and positive when I see someone who made a good bet with a bad outcome - offering them sympathies about the outcome itself of course, but also congratulating them on putting themselves out there, and making a good bet! I think it's reasonable to have some concern about insincerity or seeming mocking/insensitive, but in practice I find this often goes down well. Especially if I explain the framing of bets not outcomes, and get them to think about whether the bet was a bad idea given what they knew at the time.

This applies in all the settings I brainstormed above, but is particularly important if someone made a good bet towards me! Eg someone recommends I apply for a job that's a bad fit, sends me an article I didn't enjoy, a book recommendation I've already read, an introduction that didn't work out, gave me advice I'd already tried or

that didn't work, etc. I know I find it super discouraging to be on the other end of that, so I always try to clearly say that what they did was positive expected value, and that I appreciate it and hope they do that kind of thing again! A lot of great things in my life have come from people sending me good opportunities, and it's crazy to train people to not do that. (Though only if I think it actually *was* positive expected value, obviously - don't reward people for bad bets and bad outcomes!)

Doing this also selfishly helps me - it creates a social context around me where other people will reward me for taking good bets, and helps build the association in my mind that eg 'applying for jobs and getting rejected = good', which helps me internalise it and apply this to myself.

Exercise: Set a 5 minute timer and brainstorm ways you can help reward people around you for making good bets with bad outcomes.

Conclusion

If you relate with the failure mode of fixating on failures and being risk-averse, I think it's *really* worth trying to be on top of this, and focusing instead on the actions you took, given what you knew at the time! Anecdotally this seems super common - many of the smartest people I know are super insecure and risk averse. And this is a massive tragedy because [the world is full of wasted motion](#) - if you're unable to be ambitious and take the opportunities that come your way, you'll miss out on a lot.

So, as a final exercise, reflect on where this bias holds you back in your own life. What are the good bets you fail to make? What opportunities do you miss out on? Where do your anxieties unduly punish you? And what are you going to do about it?