

Best of LessWrong: November 2021

1. [How To Get Into Independent Research On Alignment/Agency](#)
2. [Discussion with Eliezer Yudkowsky on AGI interventions](#)
3. [EfficientZero: How It Works](#)
4. [Feature Selection](#)
5. [Omicron Variant Post #1: We're F***ed, It's Never Over](#)
6. [larger language models may disappoint you \[or, an eternally unfinished draft\]](#)
7. [A Brief Introduction to Container Logistics](#)
8. [Ngo and Yudkowsky on alignment difficulty](#)
9. [Study Guide](#)
10. [Attempted Gears Analysis of AGI Intervention Discussion With Eliezer](#)
11. [Concentration of Force](#)
12. [Yudkowsky and Christiano discuss "Takeoff Speeds"](#)
13. [Almost everyone should be less afraid of lawsuits](#)
14. [The Rationalists of the 1950s \(and before\) also called themselves "Rationalists"](#)
15. [Preprint is out! 100,000 lumens to treat seasonal affective disorder](#)
16. [Split and Commit](#)
17. [Comments on Carlsmith's "Is power-seeking AI an existential risk?"](#)
18. [EfficientZero: human ALE sample-efficiency w/MuZero+self-supervised](#)
19. [Omicron Variant Post #2](#)
20. [The bonds of family and community: Poverty and cruelty among Russian peasants in the late 19th century](#)
21. [Ngo and Yudkowsky on AI capability gains](#)
22. [App and book recommendations for people who want to be happier and more productive](#)
23. [Soares, Tallinn, and Yudkowsky discuss AGI cognition](#)
24. [Why I'm excited about Redwood Research's current project](#)
25. [Christiano, Cotta, and Yudkowsky on AI progress](#)
26. [You are probably underestimating how good self-love can be](#)
27. [Transcript: "You Should Read HPMOR"](#)
28. [Where did the 5 micron number come from? Nowhere good. \[Wired.com\]](#)
29. [Money Stuff](#)
30. [How do we become confident in the safety of a machine learning system?](#)
31. [The Maker of MIND](#)
32. [\[Book Review\] "Sorceror's Apprentice" by Tahir Shah](#)
33. [Effective Evil](#)
34. [Comments on OpenPhil's Interpretability RFP](#)
35. [A Bayesian Aggregation Paradox](#)
36. [Coordination Skills I Wish I Had For the Pandemic](#)
37. [A positive case for how we might succeed at prosaic AI alignment](#)
38. [Covid 11/25: Another Thanksgiving](#)
39. [Why I am no longer driven](#)
40. [Worst Commonsense Concepts?](#)
41. [What would we do if alignment were futile?](#)
42. [What exactly is GPT-3's base objective?](#)
43. [Investigating Fabrication](#)
44. [Chris Voss negotiation MasterClass: review](#)
45. [Chu are you?](#)
46. [AI Tracker: monitoring current and near-future risks from superscale models](#)
47. [Why Study Physics?](#)
48. [Satisficers Tend To Seek Power: Instrumental Convergence Via Retargetability](#)
49. [Competence/Confidence](#)
50. [Paxlovid Remains Illegal: 11/24 Update](#)

Best of LessWrong: November 2021

1. [How To Get Into Independent Research On Alignment/Agency](#)
2. [Discussion with Eliezer Yudkowsky on AGI interventions](#)
3. [EfficientZero: How It Works](#)
4. [Feature Selection](#)
5. [Omicron Variant Post #1: We're F***ed, It's Never Over](#)
6. [larger language models may disappoint you \[or, an eternally unfinished draft\]](#)
7. [A Brief Introduction to Container Logistics](#)
8. [Ngo and Yudkowsky on alignment difficulty](#)
9. [Study Guide](#)
10. [Attempted Gears Analysis of AGI Intervention Discussion With Eliezer](#)
11. [Concentration of Force](#)
12. [Yudkowsky and Christiano discuss "Takeoff Speeds"](#)
13. [Almost everyone should be less afraid of lawsuits](#)
14. [The Rationalists of the 1950s \(and before\) also called themselves "Rationalists"](#)
15. [Preprint is out! 100,000 lumens to treat seasonal affective disorder](#)
16. [Split and Commit](#)
17. [Comments on Carlsmith's "Is power-seeking AI an existential risk?"](#)
18. [EfficientZero: human ALE sample-efficiency w/MuZero+self-supervised](#)
19. [Omicron Variant Post #2](#)
20. [The bonds of family and community: Poverty and cruelty among Russian peasants in the late 19th century](#)
21. [Ngo and Yudkowsky on AI capability gains](#)
22. [App and book recommendations for people who want to be happier and more productive](#)
23. [Soares, Tallinn, and Yudkowsky discuss AGI cognition](#)
24. [Why I'm excited about Redwood Research's current project](#)
25. [Christiano, Cotra, and Yudkowsky on AI progress](#)
26. [You are probably underestimating how good self-love can be](#)
27. [Transcript: "You Should Read HPMOR"](#)
28. [Where did the 5 micron number come from? Nowhere good. \[Wired.com\]](#)
29. [Money Stuff](#)
30. [How do we become confident in the safety of a machine learning system?](#)
31. [The Maker of MIND](#)
32. [\[Book Review\] "Sorceror's Apprentice" by Tahir Shah](#)
33. [Effective Evil](#)
34. [Comments on OpenPhil's Interpretability RFP](#)
35. [A Bayesian Aggregation Paradox](#)
36. [Coordination Skills I Wish I Had For the Pandemic](#)
37. [A positive case for how we might succeed at prosaic AI alignment](#)
38. [Covid 11/25: Another Thanksgiving](#)
39. [Why I am no longer driven](#)
40. [Worst Commonsense Concepts?](#)
41. [What would we do if alignment were futile?](#)
42. [What exactly is GPT-3's base objective?](#)
43. [Investigating Fabrication](#)
44. [Chris Voss negotiation MasterClass: review](#)
45. [Chu are you?](#)
46. [AI Tracker: monitoring current and near-future risks from superscale models](#)
47. [Why Study Physics?](#)

48. [Satisficers Tend To Seek Power: Instrumental Convergence Via Retargetability](#)
49. [Competence/Confidence](#)
50. [Paxlovid Remains Illegal: 11/24 Update](#)

How To Get Into Independent Research On Alignment/Agency

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

I'm an independent researcher working on AI alignment and the theory of agency. I'm 29 years old, will make about \$90k this year, and set my own research agenda. I deal with basically zero academic bullshit - my grant applications each take about one day's attention to write (and decisions typically come back in ~1 month), and I publish the bulk of my work right here on [LessWrong/AF](#). Best of all, I work on some [really cool technical problems](#) which I expect are central to the future of humanity.

If your reaction to that is "Where can I sign up?", then this post is for you.

Background Models

Independence

First things first: the "independent" part of "independent research" means self-employment, and everything that goes with it. It means the onus is on you to figure out what to do, how to provide value, what to prioritize, and what to aim for. In practice, it also usually means "independent" in a broader sense: you won't have a standard template or agenda to follow. If you go down this path, assume that **you will need to chart your own course** - in particular, your own research agenda.

For the sort of person this post is aimed at, that will be a very big *upside*, not a downside.

Disclaimer: there are ways to get into alignment research which don't involve *quite* so much figuring-it-all-out-on-your-own. Some people receive mentorship from existing researchers. Some people go work for alignment research organizations. Either of those paths can involve "independent research" in the sense that you are technically self-employed, but those paths aren't "independent" in the broader sense of the word, and they're not the main topic of this post.

Preparadigmicity

As a field, the study of alignment and agency is especially well-suited to independent research, because they center around [problems we don't understand](#). It's not just that we don't have the answers; we don't even have the right frames for thinking about the problems. Agency is an area where [we are fundamentally confused](#). AI alignment is largely [a problem which hasn't happened yet, on technology which hasn't been invented yet, which we nonetheless want to solve in advance](#). Figuring out the right frames - the right paradigm - is itself a central part of the job.

The field needs people who are going to come up with new frames/approaches/models/paradigms/etc, because we're pretty sure the current frames/approaches/models/paradigms/etc aren't enough. Thus the great fit for

independent research: as an independent researcher, you're not beholden to some existing agenda based on existing frames. Coming up with your own idea of what the key problems are, how to frame them, what tools to apply... that sort of thing is exactly what we need, and it requires people who aren't committed to the strategies of existing senior researchers and organizations. It requires people who have an independent high-level understanding of the field and different angles of looking at, and can pick out the key problems and paths from that perspective.

Again, for the sort of person this post is aimed at, that will be a very big upside.

... but it comes with some trade-offs. As a historical example of pre-paradigmatic research, here's [Kuhn](#) talking about optics before Newton:

Being able to take no common body of belief for granted, each writer on physical optics felt forced to build his field anew from its foundations. In doing so, his choice of supporting experiment and observation was relatively free, for there was no standard set of methods or of phenomena that every optical writer felt forced to employ and explain. Under these circumstances, the dialogue of the resulting books was often directed as much to the members of other schools as it was to nature.

This very much applies to alignment research. Because the field does not already have a set of [shared frames](#) - i.e. a paradigm - you will need to spend a *lot* of effort explaining your frames, tools, agenda, and strategy. For the field, such discussion is a necessary step to spreading ideas and eventually creating a paradigm. For you, it's a necessary step to get paid, and to get useful engagement with your work from others.

In particular, you will probably need to both think and write a lot about your strategy: the models and intuitions which inform *why* you're working on the particular problems you've chosen, why the tools you're using seem promising, what kinds of results you expect, and what your long-term vision looks like. Inevitably, a lot of this will rely on informal arguments or intuitions; you will need to figure out how to trace the sources of those intuitions and explain them to other people, without having to formalize everything. [Explain the actual process which led to an idea/decision/approach](#), without going down the bottomless rabbit hole of deeply researching every single claim.

The current version of LessWrong was built in large part to support exactly that sort of discussion, and I strongly recommend using it.

Getting Paid

Right now, the best grantmaker in this space is the [Long-Term Future Fund \(LTFF\)](#). There are [other options](#), but none are quite as good a fit for the sort of work we're talking about here.

I've received a few LTFF grants myself and know some of the people involved in the grantmaking decisions, so I'll give some thoughts on the most important things you'll need in order to get paid. Bear in mind that this is inherently speculative and not endorsed by anyone at LTFF. I'd also recommend looking at LTFF's [past grants](#) to get a more direct idea of what kinds of things they fund.

Don't Bullshit

A low-bullshit grantmaking process works both ways. The LTFF wants to do object-level useful things, not just Look Prestigious, so they keep the application simple and the turnaround time relatively fast. The flip side is that I expect them to look very unkindly on bullshit - i.e. attempts to make the applicant/application Sound Prestigious without actually doing object-level useful things.

In academia, it's common practice to make up some bullshit about how your research is going to help the world. During my undergrad, this sort of bullshit was explicitly taught. Of course, it's not like anyone is ever going to hire an economist or statistician (let alone consult a prediction market) to figure out whether the research is *actually* likely to impact the world in the manner claimed. The goal is just to make the proposal sound good. If you're coming from academia, this sort of bullshit may be an ingrained habit which takes effort to break.

If you want to make it in alignment/agency research, you're going to need an actual object-level strategy.

We'll talk more in the next sections about how to come up with a strategy, but the first stop is [The Bottom Line](#): once you've chosen a strategy, anything you say to justify it will not make it any more correct. All that matters is the process which *originally* made you choose that strategy, or made you stick to it at times when you might realistically have changed course. So first things first, forget whatever clever idea you already have cached, and let's start from a blank slate.

Reading

Preparadigmicity means you'll need to spend a lot of time explaining your choice of vision, strategy, models, tools, etc. The flip side of that coin is reading: you'll probably need to read quite a bit of material from others in the field. This is often nontechnical or semi-technical background material, explanations of intuitions, vague gesturing at broad ideas, etc - you can see plenty of it here on LessWrong and the Alignment Forum. The more of this you read, the better you'll understand other researchers' frames (or at least know *which* frames you *don't* understand), and the better you'll be able to explain your own material in terms others can readily understand.

Early on, there are two main motivators for reading:

- To understand which strategies have already been tried, and failed, to avoid retreading that ground
- To understand a bit of the existing jargon (definitely not all of it!), in order to explain your own ideas in terms already familiar to others

To understand (some) existing approaches and jargon, I'd recommend at least skimming these sequences/posts, and diving deeper into whichever most resemble the directions you want to pursue:

- [Embedded Agency](#)
- [Value Learning](#)
- [11 Proposals For Building Safe Advanced AI](#)
- [Risks From Learned Optimization](#)

To understand barriers (other than what's discussed in the above links), [this talk](#) and the [Rocket Alignment Problem](#) are probably the best starting points. Note that lots of

people disagree with those last two links (as well as 11 Proposals), but you probably want to be at least familiar enough to have an *informed* disagreement.

Note that this is all on LessWrong, which means you can leave comments with questions, attempts to summarize, disagreements, etc. Often people will reply. This helps a lot for actually absorbing the ideas. (h/t Adam Shimi for pointing this out.)

I invite others to leave suggested reading in the comments. (This does risk turning into a big debate over whether X or Y is actually a good idea for new people, but at least then we'll have a realistic demonstration of how much everybody disagrees over all this. I did warn you that the field is preparadigmatic!)

Finally, there's [The Sequences](#). They are long, but if you haven't read them, then you definitely risk various failure modes which will be obvious to people who have read them and very confusing to you. I wouldn't quite say they're required reading, especially if you're on the more technical end of the spectrum and already somewhat familiar with alignment discussions, but there are definitely many people who will be somewhat surprised if you do technical alignment/agency research and haven't read them.

Again, I want to emphasize that everyone disagrees on all this stuff. Roughly speaking, assume that the grantmakers care more about your research having *some* plausible path to usefulness than about agreeing with any particular position in any of the field's ongoing arguments.

The Hamming Question

Over on the other side of the dining hall was a chemistry table. I had worked with one of the fellows, Dave McCall; furthermore he was courting our secretary at the time. I went over and said, "Do you mind if I join you?" They can't say no, so I started eating with them for a while. And I started asking, "What are the important problems of your field?" And after a week or so, "What important problems are you working on?" And after some more time I came in one day and said, "If what you are doing is not important, and if you don't think it is going to lead to something important, why are you at Bell Labs working on it?" I wasn't welcomed after that; I had to find somebody else to eat with!

Probably the most common mistake people make when first attempting to enter the alignment/agency research field is to not have any model at all of the main bottlenecks to alignment, or how their work will address those bottlenecks. The standard (and strongly recommended) exercise to alleviate that problem is to start from the [Hamming Questions](#):

- What are the most important problems in your field (i.e. alignment/agency)?
- How are you going to solve them?

At this point, somebody usually complains that minor contributions are important or some such. I'm not going to argue with that, because I expect the sort of person who this post is already aimed at (i.e. people who are excited to forge their own path in a technical field where everyone is fundamentally confused) is probably not the sort of person who is aiming for minor contributions anyway.

If you have decent answers to the Hamming Questions, and you make those answers clear to other people, that is probably a sufficient condition for your grant application

to not end up in the giant pile of applications from people who don't even have a model of how their proposal will help. It's not *quite* a sufficient condition to get paid, but I would guess that a large majority of people who can clearly answer the Hamming Questions do get paid.

I want to emphasize that I think clear answers to the Hamming Questions are an approximately-sufficient condition, not an approximately-necessary condition; there are definitely other paths. [Steve's story](#) in the comments below is a good example; in his words:

If you're a kinda imposter-syndrome-y person who just constitutionally wouldn't dream of looking themselves in the mirror and saying "I am aiming for a major contribution!", well me too, and don't let John scare you off. :-P

Use Your Pareto Frontier

A great line from Adam Shimi:

Most people who try to go in a direction 'no one else has tried' end up going in the most obvious direction which everyone else has tried.

My main advice to avoid this failure mode is to [leverage your Pareto frontier](#). Apply whatever knowledge, or combination of knowledge, you have which others in the field don't. Personally, I've gained a lot of insight into agency by drawing on systems biology, economics, statistical mechanics, and chaos theory. Others draw heavily on abstract math, like category theory or model theory. Evolutionary biology and user interface design are both rich sources.

This is one reason why it helps to have a broad technical background: the more frames and tools you have to draw on, the more likely you'll find a novel and promising combination to apply to the most important problems in the field. (Or, just as good: the more frames and tools you have to draw on, the more likely you'll notice that one of the most important problems has been overlooked.)

Flip side of this: if you have a novel-seeming idea which involves the same kinds of frames and tools which most people in alignment have (i.e. programming expertise, some ML experience, reading [Astral Codex Ten](#)) then do write it up, but don't be surprised if it's already been done.

If you read through some existing alignment work, and the strategy seems obviously wrong to you in a way which would not be obvious to the median LessWrong user, then that's a very promising sign.

Legibility

Part of getting a grant is not just having a good plan and the skills to execute it, but to make your plan and skills legible to the people reviewing the grant.

Here's (my summary of) a rough model from Oli, who's one of the fund managers for LTFF. In order to get a grant for alignment research, usually someone needs to do *one* of these three:

1. Write a grant application which clearly signals that they understand the alignment problem and have a non-bullshitted research strategy. (This is rare/difficult.)
2. Have a reference from someone the fund managers know and trust (i.e. the existing alignment research community).
3. Have some visible online material which clearly signals that they understand the alignment problem and have a non-bullshitted research strategy. (LessWrong posts/comments are a central example.)

As a new entrant to the field, I expect that option #3 is probably your main path. Write up not just your research strategy, but the intuitions, models and arguments behind that strategy. Give examples. Explain what you consider the key problems, why those problems seem central, and the frames and generators behind that reasoning. Again, give examples. Explain conjectures or tools you think are relevant, ideally with examples. If you're on the theory side, sketch potential empirical tests; if on the empirical side, sketch the conceptual theory behind the ideas. And include examples. Explain your vision of success, and expected applications of your research (if it succeeds). At all stages, focus on giving accessible, intuitive explanations and **lots of examples**; even people who have lots of technical background will often skip over sections with just dense math, and not everyone has the *same* technical background as you. And [put the examples at the beginnings of the posts, before the abstract/general explanations.](#)

Remember: this is preparadigmatic work. Writing up the ideas, and the generators of the ideas, and the frames, and the tools, and making it all clear and accessible to people with totally different frames and tools, is a central part of the job.

All this writing will also make option #1 and #2 easier over time: writing a lot of posts and comments will eventually generate social connections (though this takes quite a bit of time, especially if you're not in the Bay Area), and discussion/feedback will give some idea of how to explain things in a way which signals the kinds-of-things LTFF looks for.

(On the topic of feedback: a lot of more experienced researchers ignore most posts which they don't find very promising, partly because it's a lot of work to explain/argue about problems and partly because there are too many posts to read it all anyway. If you explicitly reach out - e.g. send a message on LessWrong - and ask for feedback, people are much more likely to tell you what they think.)

By the time all that is written up and posted, the grant application itself is a drop in the bucket; that's a big part of why it only takes a day to write up. A quote from Oli regarding the actual application:

I really wish people would just pretend they're writing me an email explaining what they plan to do, rather than something aimed at the general public.

This is part of why option #1 is rare - people try to write the LTFF application like it's an academic grant application or something, and it really isn't. But also, clear communication is just pretty hard in general, even when you do understand the problem and have a non-bullshitted strategy.

When To Start

This post was mostly written for people who already have the technical skills they need. That probably means grad-level education, though a PhD is definitely not a formal requirement. I know at least a few who think less-than-a-full-undergrad can suffice. Personally, I never went to grad school (though admittedly my undergrad coursework looks an awful lot like a PhD program; I got an unusually large amount of mileage out of it).

In terms of specific skills, I recently wrote a [study guide](#) with a bunch of technical topics I've found useful, but the more important point is that we don't currently know what the right combination of background knowledge is. If you already have a broad technical background, then my advice is to take a stab at the problem and see how it goes.

If you are currently in high school or undergrad, the [study guide](#) has some recommendations for what to study (and why). The larger your knowledge base, the more tools and frames you'll have to draw on later. You could also apply for a grant to e.g. pursue some alignment/agency research project over the summer; taking a stab at it will give you some firsthand data on what kinds of tools/frames are useful.

Runway

The grant application takes maybe a day, but there will probably be some groundwork before you're ready for that. You'll probably want to read a bunch, figure out a strategy, put up a few posts on it, and maybe update in response to feedback.

Personally, I quit my job as a data scientist in late 2018, and tried out a few different things over the course of the next year before settling into alignment/agency research. I got my first grant in late 2019. If someone with roughly my 2018 level of background knew up front that they wanted to enter the field, I think it would take a lot less time than that; a few months would be my guess. That said, my level of background in 2018 was already well above zero.

I wrote a fair bit on LessWrong, and researched some agency problems, even before quitting my job. I do expect it helps to "ease into it" this way, and if you're coming in fresh you should probably give yourself extra time to start writing up ideas, following the field, and getting feedback. That said, you should probably plan on going full time *at latest* by the time you get a grant, and possibly sooner. If you're in academia, then you'll probably have more room to aim the bulk of your research at alignment without striking out on your own. (Though you should still totally strike out on your own and enjoy the no-academic-bullshit lifestyle.)

Meta

Historically, EA causes (including alignment) have largely drawn from very young populations (mostly undergrads). I believe this is mostly because (a) those are the people who don't need to be drawn away from a different path which they're already on, (b) they're willing to work for peanuts, and (c) they don't have to unlearn how to bullshit. Unfortunately, a lot of alignment research benefits from a broad technical background, which takes time to build up. So I think we've historically had fewer researchers with that sort of broad knowledge than would be ideal, just because we tend to recruit young people.

But conditions have changed in recent years, and I think there's now room for a different kind of recruitment, aimed at (somewhat) older people with more knowledge and experience.

First: the Sequences are about ten years old, so right about now there are probably a bunch of postgrads and adjunct professors with lots of technical skills who have already read them, have decent epistemic habits (i.e. know how to *not bullshit*), and have a rough understanding of what the alignment problem is.

Second: nowadays, we have money. If you're a postgrad or adjunct professor or whatever, and you can do good technical alignment research, you can probably make *more* money as an independent researcher in alignment than you do now. Our [main grantmaker](#) has an application form which takes maybe a few hours at most, usually comes back with a decision in under a month, and complains that it doesn't have enough good projects to spend its money on.

So if you're the sort of person who:

- Wants to tackle big open research problems
- ... in a field where everyone is confused and we don't have a paradigm yet and you have to basically chart your own course
- ... and the stakes are literally astronomical
- ... and you have a bunch of technical skills, maybe read the sequences ten years ago, and have a basic understanding of what AI alignment is and why it's hard

... then now is a good time to sit down with a notebook and think about how you'd go about understanding alignment/agency. If you have any promising ideas, write them up, post them here on LessWrong, and apply for a grant to pursue this research full-time.

I can attest that it's an awesome job.

Discussion with Eliezer Yudkowsky on AGI interventions

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

The following is a partially redacted and lightly edited transcript of a chat conversation about AGI between Eliezer Yudkowsky and a set of invitees in early September 2021. By default, all other participants are anonymized as "Anonymous".

I think this Nate Soares quote (excerpted from Nate's [response to a report by Joe Carlsmith](#)) is a useful context-setting preface regarding timelines, which weren't discussed as much in the transcript:

[...] My odds [of AGI by the year 2070] are around 85%[...]

I can list a handful of things that drive my probability of AGI-in-the-next-49-years above 80%:

1. 50 years ago was 1970. The gap between AI systems then and AI systems now seems pretty plausibly greater than the remaining gap, even before accounting the recent dramatic increase in the rate of progress, and potential future increases in rate-of-progress as it starts to feel within-grasp.

2. I observe that, 15 years ago, everyone was saying AGI is far off because of what it couldn't do -- basic image recognition, go, starcraft, winograd schemas, programmer assistance. But basically all that has fallen. The gap between us and AGI is made mostly of intangibles. (Computer Programming That Is Actually Good? Theorem proving? Sure, but on my model, "good" versions of those are a hair's breadth away from full AGI already. And the fact that I need to clarify that "bad" versions don't count, speaks to my point that the only barriers people can name right now are intangibles.) That's a very uncomfortable place to be!

3. When I look at the history of invention, and the various anecdotes about the Wright brothers and Enrico Fermi, I get an impression that, when a technology is pretty close, the world looks a lot like how our world looks.

- Of course, the trick is that when a technology is a little far, the world might also look pretty similar!
- Though when a technology is **very** far, the world **does** look different -- it looks like experts pointing to specific technical hurdles. We exited that regime a few years ago.

4. Summarizing the above two points, I suspect that I'm in more-or-less the "penultimate epistemic state" on AGI timelines: I don't know of a project that seems like they're right on the brink; that would put me in the "final epistemic state" of thinking AGI is imminent. But I'm in the second-to-last epistemic state, where I wouldn't feel all that shocked to learn that some group has reached the brink. Maybe I won't get that call for 10 years! Or 20! But it could also be 2, and I wouldn't get to be indignant with reality. I wouldn't get to say "but all the following things should have happened first, before I made that observation". I have made those observations.

5. It seems to me that the Cotra-style compute-based model provides pretty conservative estimates. For one thing, I don't expect to need human-level compute to get human-level intelligence, and for another I think there's a decent chance that insight and innovation have a big role to play, especially on 50 year timescales.

6. There has been a lot of AI progress recently. When I tried to adjust my beliefs so that I was **positively** surprised by AI progress just about as often as I was **negatively** surprised by AI progress, I ended up expecting a bunch of rapid progress. [...]

Further preface by Eliezer:

In some sections here, I sound gloomy about the probability that coordination between AGI groups succeeds in saving the world. Andrew Critch reminds me to point out that gloominess like this can be a self-fulfilling prophecy - if people think successful coordination is impossible, they won't try to coordinate. I therefore remark in retrospective advance that it seems to me like at least some of the top AGI people, say at Deepmind and Anthropic, are the sorts who I think would rather coordinate than destroy the world; my gloominess is about what happens when the technology has propagated further than that. But even then, anybody who would *rather* coordinate and *not* destroy the world shouldn't rule out hooking up with Demis, or whoever else is in front if that person also seems to prefer not to completely destroy the world. (Don't be too picky here.) Even if the technology proliferates and the world ends a year later when other non-coordinating parties jump in, it's still better to take the route where the world ends one year later instead of immediately. Maybe the horse will sing.

Eliezer Yudkowsky

Hi and welcome. Points to keep in mind:

- I'm doing this because I would like to learn whichever *actual* thoughts this target group may have, and perhaps respond to those; that's part of the point of anonymity. If you speak an anonymous thought, please have that be your actual thought that you are thinking yourself, not something where you're thinking "well, somebody else might think that..." or "I wonder what Eliezer's response would be to..."
- Eliezer's responses are uncloaked by default. Everyone else's responses are anonymous (not pseudonymous) and neither I nor MIRI will know which potential invitee sent them.
- Please do not reshare or pass on the link you used to get here.
- I do intend that parts of this conversation may be saved and published at MIRI's discretion, though not with any mention of who the anonymous speakers could possibly have been.

Eliezer Yudkowsky

(Thank you to Ben Weinstein-Raun for building chathamroom.com, and for quickly adding some features to it at my request.)

Eliezer Yudkowsky

It is now 2PM; this room is now open for questions.

Anonymous

How long will it be open for?

Eliezer Yudkowsky

In principle, I could always stop by a couple of days later and answer any unanswered questions, but my basic theory had been "until I got tired".

Anonymous

At a high level one thing I want to ask about is research directions and prioritization. For example, if you were dictator for what researchers here (or within our influence) were working on, how would you reallocate them?

Eliezer Yudkowsky

The first reply that came to mind is "I don't know." I consider the present gameboard to look incredibly grim, and I don't actually see a way out through hard work alone. We can hope there's a miracle that violates some aspect of my background model, and we can try to prepare for that unknown miracle; preparing for an unknown miracle probably looks like "Trying to die with more dignity on the mainline" (because if you can die with more dignity on the mainline, you are better positioned to take advantage of a miracle if it occurs).

Anonymous

I'm curious if the grim outlook is currently mainly due to technical difficulties or social/coordination difficulties. (Both avenues might have solutions, but maybe one seems more recalcitrant than the other?)

Eliezer Yudkowsky

Technical difficulties. Even if the social situation were vastly improved, on my read of things, everybody still dies because there is nothing that a handful of socially coordinated projects can do, or even a handful of major governments who aren't

willing to start nuclear wars over things, to prevent somebody else from building AGI and killing everyone 3 months or 2 years later. There's no obvious winnable position into which to play the board.

Anonymous

just to clarify, that sounds like a large scale coordination difficulty to me (i.e., we - as all of humanity - can't coordinate to not build that AGI).

Eliezer Yudkowsky

I wasn't really considering the counterfactual where humanity had a collective telepathic hivemind? I mean, I've written fiction about a world coordinated enough that they managed to shut down all progress in their computing industry and only manufacture powerful computers in a single worldwide hidden base, but Earth was never going to go down that route. Relative to remotely plausible levels of future coordination, we have a technical problem.

Anonymous

Curious about why building an AGI aligned to its users' interests isn't a thing a handful of coordinated projects could do that would effectively prevent the catastrophe. The two obvious options are: it's too hard to build it vs it wouldn't stop the other group anyway. For "it wouldn't stop them", two lines of reply are nobody actually wants an unaligned AGI (they just don't foresee the consequences and are pursuing the benefits from automated intelligence, so can be defused by providing the latter) (maybe not entirely true: omnicidal maniacs), and an aligned AGI could help in stopping them. Is your take more on the "too hard to build" side?

Eliezer Yudkowsky

Because it's too technically hard to align some cognitive process that is powerful enough, and operating in a sufficiently dangerous domain, to stop the next group from building an unaligned AGI in 3 months or 2 years. Like, they can't coordinate to build an AGI that builds a nanosystem because it is too technically hard to align their AGI technology in the 2 years before the world ends.

Anonymous

Summarizing the threat model here (correct if wrong): The nearest competitor for building an AGI is at most N (<2) years behind, and building an aligned AGI, even when starting with the ability to build an unaligned AGI, takes longer than N years. So at some point some competitor who doesn't care about safety builds the unaligned AGI. How does "nobody actually wants an unaligned AGI" fail here? It takes >N years

to get everyone to realise that they have that preference and that it's incompatible with their actions?

Eliezer Yudkowsky

Many of the current actors seem like they'd be really gung-ho to build an "unaligned" AGI because they think it'd be super neat, or they think it'd be super profitable, and they don't expect it to destroy the world. So if this happens in anything like the current world - and I neither expect vast improvements, nor have very long timelines - then we'd see Deepmind get it first; and, if the code was not *immediately* stolen and rerun with higher bounds on the for loops, by China or France or whoever, somebody else would get it in another year; if that somebody else was Anthropic, I could maybe see them also not amping up their AGI; but then in 2 years it starts to go to Facebook AI Research and home hobbyists and intelligence agencies stealing copies of the code from other intelligence agencies and I don't see how the world fails to end past that point.

Anonymous

What does trying to die with more dignity on the mainline look like? There's a real question of prioritisation here between solving the alignment problem (and various approaches within that), and preventing or slowing down the next competitor. I'd personally love more direction on where to focus my efforts (obviously you can only say things generic to the group).

Eliezer Yudkowsky

I don't know how to effectively prevent or slow down the "next competitor" for more than a couple of years even in plausible-best-case scenarios. Maybe some of the natsec people can be grownups in the room and explain why "stealing AGI code and running it" is as bad as "full nuclear launch" to their foreign counterparts in a realistic way. Maybe more current AGI groups can be persuaded to go closed; or, if more than one has an AGI, to coordinate with each other and not rush into an arms race. I'm not sure I believe these things can be done in real life, but it seems understandable to me how I'd go about trying - though, please do talk with me a lot more before trying anything like this, because it's easy for me to see how attempts could backfire, it's not clear to me that we should be inviting more attention from natsec folks at all. None of that saves us without technical alignment progress. But what are other people supposed to do about researching alignment when I'm not sure what to try there myself?

Anonymous

thanks! on researching alignment, you might have better meta ideas (how to do research generally) even if you're also stuck on object level. and you might know/foresee dead ends that others don't.

Eliezer Yudkowsky

I definitely foresee a whole lot of dead ends that others don't, yes.

Anonymous

Does pushing for a lot of public fear about this kind of research, that makes all projects hard, seem hopeless?

Eliezer Yudkowsky

What does it buy us? 3 months of delay at the cost of a tremendous amount of goodwill? 2 years of delay? What's that delay for, if we all die at the end? Even if we then got a technical miracle, would it end up impossible to run a project that could make use of an alignment miracle, because everybody was afraid of that project? Wouldn't that fear tend to be channeled into "ah, yes, it must be a government project, they're the good guys" and then the government is much more hopeless and much harder to improve upon than Deepmind?

Anonymous

I imagine lack of public support for genetic manipulation of humans has slowed that research by more than three months

Anonymous

'would it end up impossible to run a project that could make use of an alignment miracle, because everybody was afraid of that project?'

...like, maybe, but not with near 100% chance?

Eliezer Yudkowsky

I don't want to sound like I'm dismissing the whole strategy, but it sounds a *lot* like the kind of thing that backfires because you did not get *exactly* the public reaction you wanted, and the public reaction you actually got was bad; and it doesn't sound like that whole strategy actually has a visualized victorious endgame, which makes it hard to work out what the exact strategy should be; it seems more like the kind of thing that falls under the syllogism "something must be done, this is something, therefore this must be done" than like a plan that ends with humane life victorious.

Regarding genetic manipulation of humans, I think the public started out very unfavorable to that, had a reaction that was not at all exact or channeled, does not allow for any 'good' forms of human genetic manipulation regardless of

circumstances, driving the science into other countries - it is not a case in point of the intelligentsia being able to successfully cunningly manipulate the fear of the masses to some supposed good end, to put it mildly, so I'd be worried about deriving that generalization from it. The reaction may more be that the fear of the public is a big powerful uncontrollable thing that doesn't move in the smart direction - maybe the public fear of AI gets channeled by opportunistic government officials into "and that's why We must have Our AGI first so it will be Good and we can Win". That seems to me much more like a thing that would happen in real life than "and then we managed to manipulate public panic down exactly the direction we wanted to fit into our clever master scheme", especially when we don't actually *have* the clever master scheme it fits into.

Eliezer Yudkowsky

I have a few stupid ideas I could try to investigate in ML, but that would require the ability to run significant-sized closed ML projects full of trustworthy people, which is a capability that doesn't seem to presently exist. Plausibly, this capability would be required in any world that got some positive model violation ("miracle") to take advantage of, so I would want to build that capability today. I am not sure how to go about doing that either.

Anonymous

if there's a chance this group can do something to gain this capability I'd be interested in checking it out. I'd want to know more about what "closed" and "trustworthy" mean for this (and "significant-size" I guess too). E.g., which ones does Anthropic fail?

Eliezer Yudkowsky

What I'd like to exist is a setup where I can work with people that I or somebody else has vetted as seeming okay-trustworthy, on ML projects that aren't going to be published. Anthropic looks like it's a package deal. If Anthropic were set up to let me work with 5 particular people at Anthropic on a project boxed away from the rest of the organization, that would potentially be a step towards trying such things. It's also not clear to me that Anthropic has either the time to work with me, or the interest in doing things in AI that aren't "stack more layers" or close kin to that.

Anonymous

That setup doesn't sound impossible to me -- at DeepMind or OpenAI or a new org specifically set up for it (or could be MIRI) -- the bottlenecks are access to trustworthy ML-knowledgeable people (but finding 5 in our social network doesn't seem impossible?) and access to compute (can be solved with more money - not too hard?). I don't think DM and OpenAI are publishing everything - the "not going to be published" part doesn't seem like a big barrier to me. Is infosec a major bottleneck (i.e., who's potentially stealing the code/data)?

Anonymous

Do you think Redwood Research could be a place for this?

Eliezer Yudkowsky

Maybe! I haven't ruled RR out yet. But they also haven't yet done (to my own knowledge) anything demonstrating the same kind of AI-development capabilities as even GPT-3, let alone AlphaFold 2.

Eliezer Yudkowsky

I would potentially be super interested in working with Deepminders if Deepmind set up some internal partition for "Okay, accomplished Deepmind researchers who'd rather not destroy the world are allowed to form subpartitions of this partition and have their work not be published outside the subpartition let alone Deepmind in general, though maybe you have to report on it to Demis only or something." I'd be more skeptical/worried about working with OpenAI-minus-Anthropic because the notion of "open AI" continues to sound to me like "what is the worst possible strategy for making the game board as unplayable as possible while demonizing everybody who tries a strategy that could possibly lead to the survival of humane intelligence", and now a lot of the people who knew about that part have left OpenAI for elsewhere. But, sure, if they changed their name to "ClosedAI" and fired everyone who believed in the original OpenAI mission, I would update about that.

Eliezer Yudkowsky

Context that is potentially missing here and should be included: I wish that Deepmind had more internal closed research, and internally siloed research, as part of a larger wish I have about the AI field, independently of what projects I'd want to work on myself.

The present situation can be seen as one in which a common resource, the remaining timeline until AGI shows up, is incentivized to be burned by AI researchers because they have to come up with neat publications and publish them (which burns the remaining timeline) in order to earn status and higher salaries. The more they publish along the spectrum that goes {quiet internal result -> announced and demonstrated result -> paper describing how to get the announced result -> code for the result -> model for the result}, the more timeline gets burned, and the greater the internal and external prestige accruing to the researcher.

It's futile to wish for everybody to act uniformly against their incentives. But I think it would be a step forward if the relative incentive to burn the commons could be reduced; or to put it another way, the more researchers have the *option* to not burn the timeline commons, without them getting fired or passed up for promotion, the more that unusually intelligent researchers might perhaps decide not to do that. So I wish in general that AI research groups in general, but also Deepmind in particular, would have affordances for researchers who go looking for interesting things to not publish any resulting discoveries, at all, and still be able to earn internal points for

them. I wish they had the *option* to do that. I wish people were *allowed* to not destroy the world - and still get high salaries and promotion opportunities and the ability to get corporate and ops support for playing with interesting toys; if destroying the world is prerequisite for having nice things, nearly everyone is going to contribute to destroying the world, because, like, they're not going to just *not* have nice things, that is not human nature for almost all humans.

When I visualize how the end of the world plays out, I think it involves an AGI system which has the ability to be cranked up by adding more computing resources to it; and I think there is an extended period where the system is not aligned enough that you can crank it up that far, without everyone dying. And it seems *extremely* likely that if factions on the level of, say, Facebook AI Research, start being able to deploy systems like that, then death is very automatic. If the Chinese, Russian, and French intelligence services all manage to steal a copy of the code, and China and Russia sensibly decide not to run it, and France gives it to three French corporations which I hear the French intelligence service sometimes does, then again, everybody dies. If the builders are sufficiently worried about that scenario that they push too fast too early, in fear of an arms race developing very soon if they wait, again, everybody dies.

At present we're very much waiting on a miracle for alignment to be possible at all, even if the AGI-builder successfully prevents proliferation and has 2 years in which to work. But if we get that miracle at all, it's not going to be an instant miracle. There'll be some minimum time-expense to do whatever work is required. So any time I visualize anybody trying to even start a successful trajectory of this kind, they need to be able to get a lot of work done, without the intermediate steps of AGI work being published, or demoed at all, let alone having models released. Because if you wait until the last months when it is really really obvious that the system is going to scale to AGI, in order to start closing things, almost all the prerequisites will already be out there. Then it will only take 3 more months of work for somebody else to build AGI, and then somebody else, and then somebody else; and even if the first 3 factions manage not to crank up the dial to lethal levels, the 4th party will go for it; and the world ends by default on full automatic.

If ideas are theoretically internal to "just the company", but the company has 150 people who all know, plus everybody with the "sysadmin" title having access to the code and models, then I imagine - perhaps I am mistaken - that those ideas would (a) inevitably leak outside due to some of those 150 people having cheerful conversations over a beer with outsiders present, and (b) be copied outright by people of questionable allegiances once all hell started to visibly break loose. As with anywhere that handles really sensitive data, the concept of "need to know" has to be a thing, or else everyone (and not just in that company) ends up knowing.

So, even if I got run over by a truck tomorrow, I would still very much wish that in the world that survived me, Deepmind would have lots of penalty-free affordance internally for people to not publish things, and to work in internal partitions that didn't spread their ideas to all the rest of Deepmind. Like, *actual* social and corporate support for that, not just a theoretical option you'd have to burn lots of social capital and weirdness points to opt into, and then get passed up for promotion forever after.

Anonymous

What's RR?

Anonymous

It's a new alignment org, run by Nate Thomas and ~co-run by Buck Shlegeris and Bill Zito, with maybe 4-6 other technical folks so far. My take: the premise is to create an org with ML expertise and general just-do-it competence that's trying to do all the alignment experiments that something like Paul+Ajeya+Eliezer all think are obviously valuable and wish someone would do. They expect to have a website etc in a few days; the org is a couple months old in its current form.

Anonymous

How likely really is hard takeoff? Clearly, we are touching the edges of AGI with GPT and the like. But I'm not feeling this will that easily be leveraged into very quick recursive self improvement.

Eliezer Yudkowsky

Compared to the position I was arguing in the Foom Debate with Robin, reality has proved way to the further Eliezer side of Eliezer along the Eliezer-Robin spectrum. It's been very unpleasantly surprising to me how little architectural complexity is required to start producing generalizing systems, and how fast those systems scale using More Compute. The flip side of this is that I can imagine a system being scaled up to interesting human+ levels, without "recursive self-improvement" or other of the old tricks that I thought would be necessary, and argued to Robin would make fast capability gain possible. You could have fast capability gain well before anything like a FOOM started. Which in turn makes it more plausible to me that we could hang out at interesting not-superintelligent levels of AGI capability for a while before a FOOM started. It's not clear that this helps anything, but it does seem more plausible.

Anonymous

I agree reality has not been hugging the Robin kind of scenario this far.

Anonymous

Going past human level doesn't necessarily mean going "foom".

Eliezer Yudkowsky

I do think that if you get an AGI significantly past human intelligence in all respects, it would obviously tend to FOOM. I mean, I suspect that Eliezer fooms if you give an Eliezer the ability to backup, branch, and edit himself.

Anonymous

It doesn't seem to me that an AGI significantly past human intelligence necessarily tends to FOOM.

Eliezer Yudkowsky

I think in principle we could have, for example, an AGI that was just a superintelligent engineer of proteins, and of nanosystems built by nanosystems that were built by proteins, and which was corrigible enough not to want to improve itself further; and this AGI would also be dumber than a human when it came to eg psychological manipulation, because we would have asked it not to think much about that subject. I'm doubtful that you can have an AGI that's significantly above human intelligence in *all* respects, without it having the capability-if-it-wanted-to of looking over its own code and seeing lots of potential improvements.

Anonymous

Alright, this makes sense to me, but I don't expect an AGI to *want* to manipulate humans that easily (unless designed to). Maybe a bit.

Eliezer Yudkowsky

Manipulating humans is a convergent instrumental strategy if you've accurately modeled (even at quite low resolution) what humans are and what they do in the larger scheme of things.

Anonymous

Yes, but human manipulation is also the kind of thing you need to guard against with even mildly powerful systems. Strong impulses to manipulate humans, should be vetted out.

Eliezer Yudkowsky

I think that, by default, if you trained a young AGI to expect that $2+2=5$ in some special contexts, and then scaled it up without further retraining, a generally superhuman version of that AGI would be very likely to 'realize' in some sense that $SS0+SS0=SSSS0$ was a consequence of the Peano axioms. There's a natural/convergent/coherent output of deep underlying algorithms that generate competence in some of the original domains; when those algorithms are implicitly scaled up, they seem likely to generalize better than whatever patch on those algorithms said ' $2 + 2 = 5$ '.

In the same way, suppose that you take weak domains where the AGI can't fool you, and apply some gradient descent to get the AGI to stop outputting actions of a type

that humans can detect and label as 'manipulative'. And then you scale up that AGI to a superhuman domain. I predict that deep algorithms within the AGI will go through consequentialist dances, and model humans, and output human-manipulating actions that can't be detected as manipulative by the humans, in a way that seems likely to bypass whatever earlier patch was imbued by gradient descent, because I doubt that earlier patch will generalize as well as the deep algorithms. Then you don't get to retrain in the superintelligent domain after labeling as bad an output that killed you and doing a gradient descent update on that, because the bad output killed you. (This is an attempted very fast gloss on what makes alignment difficult *in the first place*.)

Anonymous

[i appreciate this gloss - thanks]

Anonymous

"deep algorithms within it will go through consequentialist dances, and model humans, and output human-manipulating actions that can't be detected as manipulative by the humans"

This is true if it is rewarding to manipulate humans. If the humans are on the outlook for this kind of thing, it doesn't seem that easy to me.

Going through these "consequentialist dances" to me appears to presume that mistakes that should be apparent haven't been solved at simpler levels. It seems highly unlikely to me that you would have a system that appears to follow human requests and human values, and it would suddenly switch at some powerful level. I think there will be signs beforehand. Of course, if the humans are not paying attention, they might miss it. But, say, in the current milieu, I find it plausible that they will pay enough attention.

"because I doubt that earlier patch will generalize as well as the deep algorithms"

That would depend on how "deep" your earlier patch was. Yes, if you're just doing surface patches to apparent problems, this might happen. But it seems to me that useful and intelligent systems will require deep patches (or deep designs from the start) in order to be apparently useful to humans at solving complex problems enough. This is not to say that they would be perfect. But it seems quite plausible to me that they would in most cases prevent the worst outcomes.

Eliezer Yudkowsky

"If you've got a general consequence-modeling-and-searching algorithm, it seeks out ways to manipulate humans, even if there are no past instances of a random-action-generator producing manipulative behaviors that succeeded and got reinforced by gradient descent over the random-action-generator. It invents the strategy de novo by imagining the results, even if there's no instances in memory of a strategy like that having been tried before." Agree or disagree?

Anonymous

Creating strategies de novo would of course be expected of an AGI.

"If you've got a general consequence-modeling-and-searching algorithm, it seeks out ways to manipulate humans, even if there are no past instances of a random-action-generator producing manipulative behaviors that succeeded and got reinforced by gradient descent over the random-action-generator. It invents the strategy de novo by imagining the results, even if there's no instances in memory of a strategy like that having been tried before." Agree or disagree?

I think, if the AI will "seek out ways to manipulate humans", will depend on what kind of goals the AI has been designed to pursue.

Manipulating humans is definitely an instrumentally useful kind of method for an AI, for a lot of goals. But it's also counter to a lot of the things humans would direct the AI to do -- at least at a "high level". "Manipulation", such as marketing, for lower level goals, can be very congruent with higher level goals. An AI could clearly be good at manipulating humans, while not manipulating its creators or the directives of its creators.

If you are asking me to agree that the AI will generally seek out ways to manipulate the high-level goals, then I will say "no". Because it seems to me that faults of this kind in the AI design is likely to be caught by the designers earlier. (This isn't to say that this kind of fault couldn't happen.) It seems to me that manipulation of high-level goals will be one of the most apparent kind of faults of this kind of system.

Anonymous

RE: "I'm doubtful that you can have an AGI that's significantly above human intelligence in *all* respects, without it having the capability-if-it-wanted-to of looking over its own code and seeing lots of potential improvements."

It seems plausible (though unlikely) to me that this would be true in practice for the AGI we build -- but also that the potential improvements it sees would be pretty marginal. This is coming from the same intuition that current learning algorithms might already be approximately optimal.

Eliezer Yudkowsky

If you are asking me to agree that the AI will generally seek out ways to manipulate the high-level goals, then I will say "no". Because it seems to me that faults of this kind in the AI design is likely to be caught by the designers earlier.

I expect that when people are trying to stomp out convergent instrumental strategies by training at a safe dumb level of intelligence, this will not be effective at preventing convergent instrumental strategies at smart levels of intelligence; also note that at very smart levels of intelligence, "hide what you are doing" is also a convergent instrumental strategy of that substrategy.

I don't know however if I should be explaining at this point why "manipulate humans" is convergent, why "conceal that you are manipulating humans" is convergent, why you have to train in safe regimes in order to get safety in dangerous regimes (because if you try to "train" at a sufficiently unsafe level, the output of the unaligned system deceives you into labeling it incorrectly and/or kills you before you can label the outputs), or why attempts to teach corrigibility in safe regimes are unlikely to generalize well to higher levels of intelligence and unsafe regimes (qualitatively new thought processes, things being way out of training distribution, and, the hardest part to explain, corrigibility being "anti-natural" in a certain sense that makes it incredibly hard to, eg, exhibit any coherent planning behavior ("consistent utility function") which corresponds to being willing to let somebody else shut you off, without incentivizing you to actively manipulate them to shut you off).

Anonymous

My (unfinished) idea for buying time is to focus on applying AI to well-specified problems, where constraints can come primarily from the action space and additionally from process-level feedback (i.e., human feedback providers understand why actions are good before endorsing them, and reject anything weird even if it seems to work on some outcomes-based metric). This is basically a form of boxing, with application-specific boxes. I know it doesn't scale to superintelligence but I think it can potentially give us time to study and understand proto AGIs before they kill us. I'd be interested to hear devastating critiques of this that imply it isn't even worth fleshing out more and trying to pursue, if they exist.

Anonymous

(I think it's also similar to CAIS in case that's helpful.)

Eliezer Yudkowsky

There's lots of things we can do which don't solve the problem and involve us poking around with AIs having fun, while we wait for a miracle to pop out of nowhere. There's lots of things we can do with AIs which are weak enough to not be able to fool us and to not have cognitive access to any dangerous outputs, like automatically generating pictures of cats. The trouble is that nothing we can do with an AI like that (where "human feedback providers understand why actions are good before endorsing them") is powerful enough to save the world.

Eliezer Yudkowsky

In other words, if you have an aligned AGI that builds complete mature nanosystems for you, that *is* enough force to save the world; but that AGI needs to have been aligned by some method other than "humans inspect those outputs and vet them and their consequences as safe/aligned", because humans cannot accurately and unfoolably vet the consequences of DNA sequences for proteins, or of long bitstreams sent to protein-built nanofactories.

Anonymous

When you mention nanosystems, how much is this just a hypothetical superpower vs. something you actually expect to be achievable with AGI/superintelligence? If expected to be achievable, why?

Eliezer Yudkowsky

The case for nanosystems being possible, if anything, seems even more slam-dunk than the already extremely slam-dunk case for superintelligence, because we can set lower bounds on the power of nanosystems using far more specific and concrete calculations. See eg the first chapters of Drexler's *Nanosystems*, which are the first step mandatory reading for anyone who would otherwise doubt that there's plenty of room above biology and that it is possible to have artifacts the size of bacteria with much higher power densities. I have this marked down as "known lower bound" not "speculative high value", and since *Nanosystems* has been out since 1992 and subjected to attemptedly-skeptical scrutiny, without anything I found remotely persuasive turning up, I do not have a strong expectation that any new counterarguments will materialize.

If, after reading *Nanosystems*, you still don't think that a superintelligence can get to and past the *Nanosystems* level, I'm not quite sure what to say to you, since the models of superintelligences are much less concrete than the models of molecular nanotechnology.

I'm on record as early as 2008 as saying that I expected superintelligences to crack protein folding, some people disputed that and were all like "But how do you know that's solvable?" and then AlphaFold 2 came along and cracked the protein folding problem they'd been skeptical about, far below the level of superintelligence.

I can try to explain how I was mysteriously able to forecast this truth at a high level of confidence - not the exact level where it became possible, to be sure, but that superintelligence would be sufficient - despite this skepticism; I suppose I could point to prior hints, like even human brains being able to contribute suggestions to searches for good protein configurations; I could talk about how if evolutionary biology made proteins evolvable then there must be a lot of regularity in the folding space, and that this kind of regularity tends to be exploitable.

But of course, it's also, in a certain sense, very *obvious* that a superintelligence could crack protein folding, just like it was obvious years before *Nanosystems* that molecular nanomachines would in fact be possible and have much higher power densities than biology. I could say, "Because proteins are held together by van der Waals forces that are much weaker than covalent bonds," to point to a reason how you could realize that after just reading *Engines of Creation* and before *Nanosystems* existed, by way of explaining how one could possibly guess the result of the calculation in advance of building up the whole detailed model. But in reality, precisely because the possibility of molecular nanotechnology was already obvious to any sensible person just from reading *Engines of Creation*, the sort of person who wasn't convinced by *Engines of Creation* wasn't convinced by *Nanosystems* either, because they'd already demonstrated immunity to sensible arguments; an example of the general phenomenon I've elsewhere termed the Law of Continued Failure.

Similarly, the sort of person who was like "But how do you know superintelligences will be able to build nanotech?" in 2008, will probably not be persuaded by the demonstration of AlphaFold 2, because it was already clear to anyone sensible in 2008, and so anyone who can't see sensible points in 2008 probably also can't see them after they become even clearer. There are some people on the margins of sensibility who fall through and change state, but mostly people are not on the exact margins of sanity like that.

Anonymous

"If, after reading Nanosystems, you still don't think that a superintelligence can get to and past the Nanosystems level, I'm not quite sure what to say to you, since the models of superintelligences are much less concrete than the models of molecular nanotechnology."

I'm not sure if this is directed at *me* or the https://en.wikipedia.org/wiki/Generic_you, but I'm only expressing curiosity on this point, not skepticism :)

Anonymous

some form of "scalable oversight" is the naive extension of the initial boxing thing proposed above that claims to be the required alignment method -- basically, make the humans vetting the outputs smarter by providing them AI support for all well-specified (level-below)-vettable tasks.

Eliezer Yudkowsky

I haven't seen any plausible story, in any particular system design being proposed by the people who use terms about "scalable oversight", about how human-overseeable thoughts or human-inspected underlying systems, compound into very powerful human-non-overseeable outputs that are trustworthy. Fundamentally, the whole problem here is, "You're allowed to look at floating-point numbers and Python code, but how do you get from there to trustworthy nanosystem designs?" So saying "Well, we'll look at some thoughts we can understand, and then from out of a much bigger system will come a trustworthy output" doesn't answer the hard core at the center of the question. Saying that the humans will have AI support doesn't answer it either.

Anonymous

the kind of useful thing humans (assisted-humans) might be able to vet is reasoning/arguments/proofs/explanations. without having to generate neither the trustworthy nanosystem design nor the reasons it is trustworthy, we could still check them.

Eliezer Yudkowsky

If you have an untrustworthy general superintelligence generating English strings meant to be "reasoning/arguments/proofs/explanations" about eg a nanosystem design, then I would not only expect the superintelligence to be able to fool humans in the sense of arguing for things that were not true in a way that fooled the humans, I'd expect the superintelligence to be able to covertly directly hack the humans in ways that I wouldn't understand even after having been told what happened. So you must have some prior belief about the superintelligence being aligned before you dared to look at the arguments. How did you get that prior belief?

Anonymous

I think I'm not starting with a general superintelligence here to get the trustworthy nanodesigns. I'm trying to build the trustworthy nanosystems "the hard way", i.e., if we did it without ever building AIs, and then speed that up using AI for automation of things we know how to vet (including recursively). Is a crux here that you think nanosystem design requires superintelligence?

(tangent: I think this approach works even if you accidentally built a more-general or more-intelligent than necessary foundation model as long as you're only using it in boxes it can't outsmart. The better-specified the tasks you automate are, the easier it is to secure the boxes.)

Eliezer Yudkowsky

I think that China ends the world using code they stole from Deepmind that did things the easy way, and that happens 50 years of natural R&D time before you can do the equivalent of "strapping mechanical aids to a horse instead of building a car from scratch".

I also think that the speedup step in "iterated amplification and distillation" will introduce places where the fast distilled outputs of slow sequences are not true to the original slow sequences, because gradient descent is not perfect and won't be perfect and it's not clear we'll get any paradigm besides gradient descent for doing a step like that.

Anonymous

How do you feel about the safety community as a whole and the growth we've seen over the past few years?

Eliezer Yudkowsky

Very grim. I think that almost everybody is bouncing off the real hard problems at the center and doing work that is predictably not going to be useful at the superintelligent level, nor does it teach me anything I could not have said in advance of the paper being written. People like to do projects that they know will succeed and will result in a publishable paper, and that rules out all real research at step 1 of the social process.

Paul Christiano is trying to have real foundational ideas, and they're all wrong, but he's one of the few people trying to have foundational ideas at all; if we had another 10 of him, something might go right.

Chris Olah is going to get far too little done far too late. We're going to be facing down an unalignable AGI and the current state of transparency is going to be "well look at this interesting visualized pattern in the attention of the key-value matrices in layer 47" when what we need to know is "okay but was the AGI plotting to kill us or not". But Chris Olah is still trying to do work that is on a pathway to anything important at all, which makes him exceptional in the field.

Stuart Armstrong did some good work on further formalizing the shutdown problem, an example case in point of why corrigibility is hard, which so far as I know is still resisting all attempts at solution.

Various people who work or worked for MIRI came up with some actually-useful notions here and there, like Jessica Taylor's expected utility quantilization.

And then there is, so far as I can tell, a vast desert full of work that seems to me to be mostly fake or pointless or predictable.

It is very, very clear that at present rates of progress, adding that level of alignment capability as grown over the next N years, to the AGI capability that arrives after N years, results in everybody dying very quickly. Throwing more money at this problem does not obviously help because it just produces more low-quality work.

Anonymous

"doing work that is predictably not going to be really useful at the superintelligent level, nor does it teach me anything I could not have said in advance of the paper being written"

I think you're underestimating the value of solving small problems. Big problems are solved by solving many small problems. (I do agree that many academic papers do not represent much progress, however.)

Eliezer Yudkowsky

By default, I suspect you have longer timelines and a smaller estimate of total alignment difficulty, not that I put less value than you on the incremental power of solving small problems over decades. I think we're going to be staring down the gun of a completely inscrutable model that would kill us all if turned up further, with no idea how to read what goes on inside its head, and no way to train it on humanly scrutable and safe and humanly-labelable domains in a way that seems like it would align the superintelligent version, while standing on top of a whole bunch of papers about "small problems" that never got past "small problems".

Anonymous

"I think we're going to be staring down the gun of a completely inscrutable model that would kill us all if turned up further, with no idea how to read what goes on inside its head, and no way to train it on humanly scrutable and safe and humanly-labelable domains in a way that seems like it would align the superintelligent version"

This scenario seems possible to me, but not very plausible. GPT is not going to "kill us all" if turned up further. No amount of computing power (at least before AGI) would cause it to. I think this is apparent, without knowing exactly what's going on inside GPT. This isn't to say that there aren't AI systems that wouldn't. But *what kind of system would?* (A GPT combined with sensory capabilities at the level of Tesla's self-driving AI? That still seems too limited.)

Eliezer Yudkowsky

Alpha Zero scales with more computing power, I think AlphaFold 2 scales with more computing power, Mu Zero scales with more computing power. Precisely because GPT-3 doesn't scale, I'd expect an AGI to look more like Mu Zero and particularly with respect to the fact that it has some way of scaling.

Steve Omohundro

Eliezer, thanks for doing this! I just now read through the discussion and found it valuable. I agree with most of your specific points but I seem to be much more optimistic than you about a positive outcome. I'd like to try to understand why that is. I see mathematical proof as the most powerful tool for constraining intelligent systems and I see a pretty clear safe progression using that for the technical side (the social side probably will require additional strategies). Here are some of my intuitions underlying that approach, I wonder if you could identify any that you disagree with. I'm fine with your using my name (Steve Omohundro) in any discussion of these.

- 1) Nobody powerful wants to create unsafe AI but they do want to take advantage of AI capabilities.
- 2) None of the concrete well-specified valuable AI capabilities require unsafe behavior
- 3) Current simple logical systems are capable of formalizing every relevant system involved (eg. MetaMath <http://us.metamath.org/index.html> currently formalizes roughly an undergraduate math degree and includes everything needed for modeling the laws of physics, computer hardware, computer languages, formal systems, machine learning algorithms, etc.)
- 4) Mathematical proof is cheap to mechanically check (eg. MetaMath has a 500 line Python verifier which can rapidly check all of its 38K theorems)
- 5) GPT-F is a fairly early-stage transformer-based theorem prover and can already prove 56% of the MetaMath theorems. Similar systems are likely to soon be able to rapidly prove all simple true theorems (eg. that human mathematicians can prove in a day).
- 6) We can define provable limits on the behavior of AI systems that we are confident prevent dangerous behavior and yet still enable a wide range of useful behavior.

- 7) We can build automated checkers for these provable safe-AI limits.
- 8) We can build (and eventually mandate) powerful AI hardware that first verifies proven safety constraints before executing AI software
- 9) For example, AI smart compilation of programs can be formalized and doesn't require unsafe operations
- 10) For example, AI design of proteins to implement desired functions can be formalized and doesn't require unsafe operations
- 11) For example, AI design of nanosystems to achieve desired functions can be formalized and doesn't require unsafe operations.
- 12) For example, the behavior of designed nanosystems can be similarly constrained to only proven safe behaviors
- 13) And so on through the litany of early stage valuable uses for advanced AI.
- 14) I don't see any fundamental obstructions to any of these. Getting social acceptance and deployment is another issue!

Best, Steve

Eliezer Yudkowsky

Steve, are you visualizing AGI that gets developed 70 years from now under absolutely different paradigms than modern ML? I don't see being able to take anything remotely like, say, Mu Zero, and being able to prove any theorem about it which implies anything like corrigibility or the system not internally trying to harm humans. Anything in which enormous inscrutable floating-point vectors is a key component, seems like something where it would be very hard to prove any theorems about the treatment of those enormous inscrutable vectors that would correspond in the outside world to the AI not killing everybody.

Even if we somehow managed to get structures far more legible than giant vectors of floats, using some AI paradigm very different from the current one, it still seems like huge key pillars of the system would rely on non-fully-formal reasoning; even if the AI has something that you can point to as a utility function and even if that utility function's representation is made out of programmer-meaningful elements instead of giant vectors of floats, we'd still be relying on much shakier reasoning at the point where we claimed that this utility function meant something in an intuitive human-desired sense, say. And if that utility function is learned from a dataset and decoded only afterwards by the operators, that sounds even scarier. And if instead you're learning a giant inscrutable vector of floats from a dataset, gulp.

You seem to be visualizing that we prove a theorem and then get a theorem-like level of assurance that the system is safe. What kind of theorem? What the heck would it say?

I agree that it seems plausible that the good cognitive operations we want do not *in principle* require performing bad cognitive operations; the trouble, from my perspective, is that generalizing structures that do lots of good cognitive operations

will automatically produce bad cognitive operations, especially when we dump more compute into them; "you can't bring the coffee if you're dead".

So it takes a more complicated system and some feat of insight I don't presently possess, to "just" do the good cognitions, instead of doing all the cognitions that result from decompressing the thing that compressed the cognitions in the dataset - even if that original dataset only contained cognitions that looked good to us, even if that dataset actually *was* just correctly labeled data about safe actions inside a slightly dangerous domain. Humans do a lot of stuff besides maximizing inclusive genetic fitness, optimizing purely on outcomes labeled by a simple loss function doesn't get you an internal optimizer that pursues only that loss function, etc.

Anonymous

Steve's intuitions sound to me like they're pointing at the "well-specified problems" idea from an earlier thread. Essentially, only use AI in domains where unsafe actions are impossible by construction. Is this too strong a restatement of your intuitions Steve?

Steve Omohundro

Thanks for your perspective! Those sound more like social concerns than technical ones, though. I totally agree that today's AI culture is very "sloppy" and that the currently popular representations, learning algorithms, data sources, etc. aren't oriented around precise formal specification or provably guaranteed constraints. I'd love any thoughts about ways to help shift that culture toward precise and safe approaches! Technically there is no problem getting provable constraints on floating point computations, etc. The work often goes under the label "Interval Computation". It's not even very expensive, typically just a factor of 2 worse than "sloppy" computations. For some reason those approaches have tended to be more popular in Europe than in the US. Here are a couple lists of references:

<http://www.cs.utep.edu/interval-comp/>

<https://www.mat.univie.ac.at/~neum/interval.html>

I see today's dominant AI approach of mapping everything to large networks ReLU units running on hardware designed for dense matrix multiplication, trained with gradient descent on big noisy data sets as a very temporary state of affairs. I fully agree that it would be uncontrolled and dangerous scaled up in its current form! But it's really terrible in every aspect except that it makes it easy for machine learning practitioners to quickly slap something together which will actually sort of work sometimes. With all the work on AutoML, NAS, and the formal methods advances I'm hoping we leave this "sloppy" paradigm pretty quickly. Today's neural networks are terribly inefficient for inference: most weights are irrelevant for most inputs and yet current methods do computational work on each. I developed many algorithms and data structures to avoid that waste years ago (eg. "bumptrees"

<https://steveomohundro.com/scientific-contributions/>

They're also pretty terrible for learning since most weights don't need to be updated for most training examples and yet they are. Google and others are using Mixture-of-Experts to avoid some of that cost: <https://arxiv.org/abs/1701.06538>

Matrix multiply is a pretty inefficient primitive and alternatives are being explored:
<https://arxiv.org/abs/2106.10860>

Today's reinforcement learning is slow and uncontrolled, etc. All this ridiculous computational and learning waste could be eliminated with precise formal approaches which measure and optimize it precisely. I'm hopeful that that improvement in computational and learning performance may drive the shift to better controlled representations.

I see theorem proving as hugely valuable for safety in that we can easily precisely specify many important tasks and get guarantees about the behavior of the system. I'm hopeful that we will also be able to apply them to the full AGI story and encode human values, etc., but I don't think we want to bank on that at this stage. Hence, I proposed the "Safe-AI Scaffolding Strategy" where we never deploy a system without proven constraints on its behavior that give us high confidence of safety. We start extra conservative and disallow behavior that might eventually be determined to be safe. At every stage we maintain very high confidence of safety. Fast, automated theorem checking enables us to build computational and robotic infrastructure which only executes software with such proofs.

And, yes, I'm totally with you on needing to avoid the "basic AI drives"! I think we have to start in a phase where AI systems are not allowed to run rampant as uncontrolled optimizing agents! It's easy to see how to constrain limited programs (eg. theorem provers, program compilers or protein designers) to stay on particular hardware and only communicate externally in precisely constrained ways. It's similarly easy to define constrained robot behaviors (eg. for self-driving cars, etc.) The dicey area is that unconstrained agentic edge. I think we want to stay well away from that until we're very sure we know what we're doing! My optimism stems from the belief that many of the socially important things we need AI for won't require anything near that unconstrained edge. But it's tempered by the need to get the safe infrastructure into place before dangerous AIs are created.

Anonymous

As far as I know, all the work on "verifying floating-point computations" currently is way too low-level -- the specifications that are proved about the computations don't say anything about what the computations mean or are about, beyond the very local execution of some algorithm. Execution of algorithms in the real world can have very far-reaching effects that aren't modelled by their specifications.

Eliezer Yudkowsky

Yeah, what they said. How do you get from proving things about error bounds on matrix multiplications of inscrutable floating-point numbers, to saying anything about what a mind is trying to do, or not trying to do, in the external world?

Steve Omohundro

Ultimately we need to constrain behavior. You might want to ensure your robot butler won't leave the premises. To do that using formal methods, you need to have a semantic representation of the location of the robot, your premise's spatial extent, etc. It's pretty easy to formally represent that kind of physical information (it's just a more careful version of what engineers do anyway). You also have a formal model of the computational hardware and software and the program running the system.

For finite systems, any true property has a proof which can be mechanically checked but the size of that proof might be large and it might be hard to find. So we need to use encodings and properties which mesh well with the safety semantics we care about.

Formal proofs of properties of programs has progressed to where a bunch of cryptographic, compilation, and other systems can be specified and formalized. Why it's taken this long, I have no idea. The creator of any system has an argument as to why its behavior does what they think it will and why it won't do bad or dangerous things. The formalization of those arguments should be one direct short step.

Experience with formalizing mathematician's informal arguments suggest that the formal proofs are maybe 5 times longer than the informal argument. Systems with learning and statistical inference add more challenges but nothing that seems in-principle all that difficult. I'm still not completely sure how to constrain the use of language, however. I see inside of Facebook all sorts of problems due to inability to constrain language systems (eg. they just had a huge issue where a system labeled a video with a racist term). The interface between natural language semantics and formal semantics and how we deal with that for safety is something I've been thinking a lot about recently.

Steve Omohundro

Here's a nice 3 hour long tutorial about "probabilistic circuits" which is a representation of probability distributions, learning, Bayesian inference, etc. which has much better properties than most of the standard representations used in statistics, machine learning, neural nets, etc.: <https://www.youtube.com/watch?v=2RAG5-L9R70> It looks especially amenable to interpretability, formal specification, and proofs of properties.

Eliezer Yudkowsky

You're preaching to the choir there, but even if we were working with more strongly typed epistemic representations that had been inferred by some unexpected innovation of machine learning, automatic inference of those representations would lead them to be uncommented and not well-matched with human compressions of reality, nor would they match exactly against reality, which would make it very hard for any theorem about "we are optimizing against this huge uncommented machine-learned epistemic representation, to steer outcomes inside this huge machine-learned goal specification" to guarantee safety in outside reality; especially in the face of how corrigibility is unnatural and runs counter to convergence and indeed coherence; especially if we're trying to train on domains where unaligned cognition is safe, and generalize to regimes in which unaligned cognition is not safe. Even in this case, we are not nearly out of the woods, because what we can prove has a great type-gap with

that which we want to ensure is true. You can't handwave the problem of crossing that gap even if it's a solvable problem.

And that whole scenario would require some major total shift in ML paradigms.

Right now the epistemic representations are giant inscrutable vectors of floating-point numbers, and so are all the other subsystems and representations, more or less.

Prove whatever you like about that Tensorflow problem; it will make no difference to whether the AI kills you. The properties that can be proven just aren't related to safety, no matter how many times you prove an error bound on the floating-point multiplications. It wasn't floating-point error that was going to kill you in the first place.

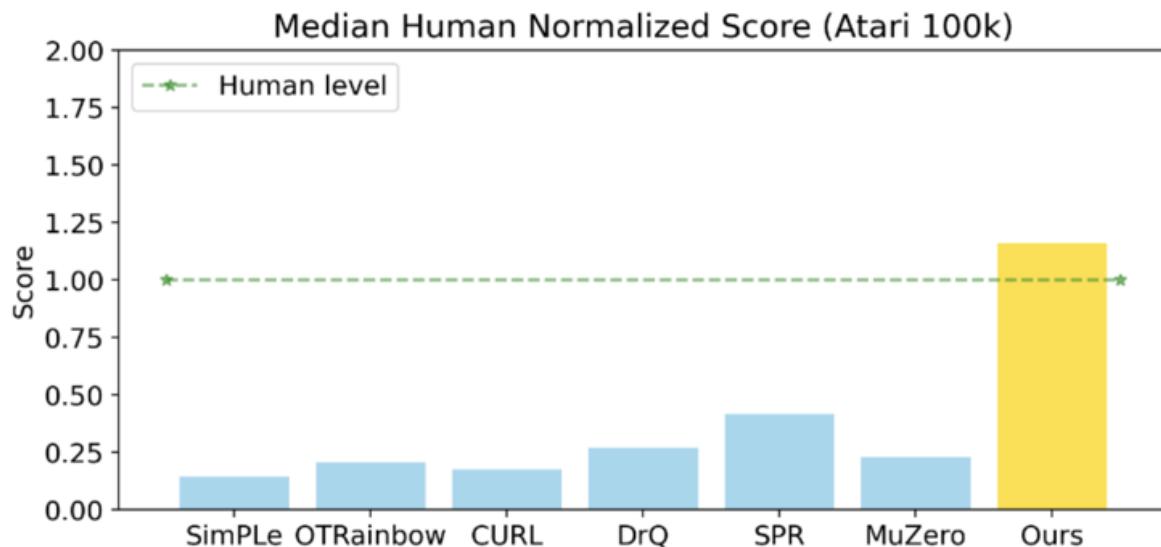
EfficientZero: How It Works

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

The goal of this essay is to help you understand [EfficientZero](#), a reinforcement learning agent that obtains better-than-human median performance on a set of 26 Atari games after just two hours of real-time experience playing each game.

Specifically, it gets 116% of human median performance on the data-limited Atari 100k benchmark. The previously-best algorithm only reached about 41% of median human performance, so this is a reasonably large leap.

Chart stolen from paper



[The benchmark](#) is called 100k because agents only interact with the environment for 100,000 steps -- about two hours. Note also that the *human* benchmarks were also set after the humans in question had about two hours of experience on the game. So EfficientZero seems to -- at least on this set of games -- exceed humans in *sample efficiency* specifically.

This is particularly impressive when you recall that, going into this, the agents were entirely ignorant of anything whatsoever about the world. The networks of their brains were initialized with random values. Humans manage to do pretty well on Atari after two hours, but we've pretrained in the actual world for many years, which lets us apply analogies from this world, experience from other video games, and so on. This agent managed to do comparably well with *none* of these advantages.

(Granted, the 100k benchmark focuses on Atari environments which are relatively easy to make progress in, because it was meant to be used for sample-efficiency benchmarks. It excludes extremely-difficult-to-explore environments like Montezuma's Revenge, where the first reward is quite hard to get.)

So. After reading this, you should understand how EfficientZero works, and how the changes in EfficientZero improve upon its predecessors. You should also know what further improvements to it are likely in the coming 3 to 24 months.

My target reader is someone who can program, or is at least broadly technically literate. They should be reasonably familiar with the basic principles of supervised machine learning, but they do not need to be particularly up-to-date with reinforcement learning. Reading this will nevertheless be heavy going if you aren't at least a little familiar with reinforcement learning; expect to have to reread some stuff.

I have aggressively and in some cases silently cut content that is not necessary for understanding the contributions of EfficientZero. For instance, in my explanation of how EfficientZero's predecessor MuZero works, I have ignored **major** elements of MuZero that EfficientZero does not modify. I have also only explained the version of MuZero ("MuZero Reanalyse") that EfficientZero builds upon. And finally in some places I'm confused about why techniques work as well as they do; I've noted them rather than try to smooth them over.

I'll divide this into five sections.

1. **Reinforcement Learning Primer:** A quick refresher on standard terminology and methods in RL, just for context. You can skip this, if you're remotely familiar with RL. What is the kind of problem that RL solves?
2. **Historical background:** EfficientZero is a relatively natural advance in model-based RL, **given** prior advances in model-free RL and elsewhere. It's useful to look at the content of these other advances, in order to understand EfficientZero and the conditions of EfficientZero's existence.
3. **MuZero:** MuZero is the 2019 algorithm which extended the Go-playing algorithm AlphaGo to single-player, non-zero sum games like Atari. How does MuZero work?
4. **EfficientZero:** EfficientZero is basically three independent modifications to MuZero stacked on top of each other. What do these three independent changes do?
5. **Conclusion.** What kind of work is likely to follow EfficientZero? Does EfficientZero provide any evidence about "how hard" it is to improve reinforcement learning right now?

1: Reinforcement Learning Primer

Alright, a quick review of the basics. If you follow RL at all, you've seen all the information below so many times at the start of various papers that it's stuck painfully to your retina. So feel free to skip.

(If you want a longer version, Sutton and Barto's "[Reinforcement Learning: An Introduction](#)" is the standard, excellent introductory text.)

The basic abstraction of a reinforcement learning problem is as follows.

There is an agent and an environment. The agent and an environment interact over a series of episodes.

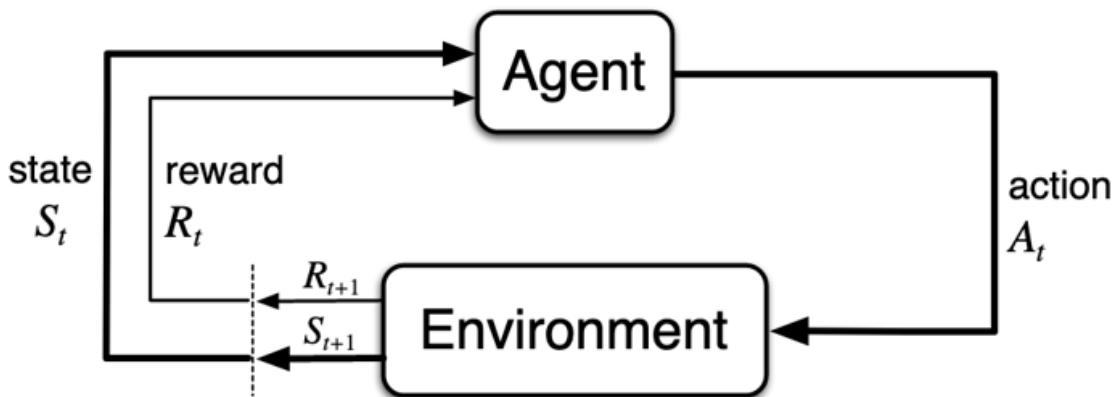
Each episode has some finite number of timesteps, counting from 0 to T. These episodes each correspond to a thing like "a game of chess" or "a level of a video game" or "an attempt to put a product in a box in a warehouse" and so on. They're always considered to be causally isolated from each other, except for the fact that an agent can learn from one to the next.

At every timestep t within such an episode, the environment emits some **observation** and **reward** to the agent, o_t and r_t . The observation is what the agent perceives; it's what the agent sees, hears, or feels; we usually represent it with a big matrix or tensor of numbers. The reward, on the other hand, is a single number that's either positive or negative, corresponding roughly to pleasure or pain.

(The observations usually need to be rolled up into a **state** s_t that contains all the relevant information for acting. This is often produced by stacking the last 4 to 10 observations — frames of video in a video game, for instance — on top of each other and saying "This is probably enough." Sometimes, the state is instead produced by a sequence-to-fixed-length-vector neural network.)

In response to the observation / state, the agent emits some **action**, a_t . The environment processes the action, and emits the next observation and reward, o_{t+1} and r_{t+1} . And so on until the episode ends.

Image from aforementioned Sutton and Barto; every RL article is contractually obligated to have some version of this image in it



Positive reward is good, obviously, and higher positive reward is better.

The agent doesn't want to maximize the immediate *per-frame* reward, though. If you chose an action which gets you 100 reward in the next frame, but 0 reward for each of a thousand frames afterwards, that's obviously inferior to an action which would get you 0 reward in the next frame, but 100 reward for each of a thousand frames afterwards. The marshmallow now might be inferior to the two marshmallows later. Maybe.

So instead the agent wants to maximize the sum of future expected reward, which is called the **return**, which I'll abbreviate G because otherwise "return" and "reward" will get confused and also because everyone else does it. This is the total of $r_t + r_{t+1} + r_{t+2} + r_{t+3} + r_{t+4} \dots$ until the end of the episode.

(The rewards that sum to make the return G are normally *discounted* according to how far in the future they are.

But in my first draft of this essay, the time-discount variable just cluttered up many subsequent formulas, which I am trying to keep simple; also, it provided no EfficientZero-specific insight. So I'll falsely pretend for the rest of the essay that EfficientZero uses an undiscounted reward, and present the math accordingly -- including in my definitions for state-value and action-value functions, which is honestly a pretty sketchy decision on my part.)

Ahem. The total return mentioned above, $r_0 + r_1 + r_2 + r_3 + \dots r_T$ can be more compactly defined using a sum notation.

$$\text{Return} = G = \sum_{t=0}^T r_t$$

Return, then, is what the RL agent wishes to maximize.

In order to do this, the agent must have a **policy**. The **policy** of an agent is just whatever makes decisions for it.

You can think of an agent's policy as a mapping from a state to an action, or from a history of observations to an action, or from a state to an action distribution. It is usually represented by the symbol π . The internals of the policy could be anything: it could include a learned model of the dynamics of the environment, it might use the state-value function defined below, it could be entirely random, and so on.

For all of what follows, I'll assume that the agent is taking discrete actions; i.e., it will choose one element out of a finite, predefined set of possible actions to perform. For instance, in a side-scrolling video game, the set might include go left, go right, jump, and attack.

Specifically, we'll represent the output of a policy as a *distribution over actions*. Thus, if an agent could do one of go left, go right, or be still, a policy might output [0.33, 0.33, 0.33] to indicate that it is ambivalent between the actions, or [0.9, 0.05, 0.05] to indicate it has a strong preference for one or another. Then when actually acting, the agent samples randomly from this policy distribution.

Because of this, you can think of a policy as a function taking a state and an action, and returning a probability -- $\pi(s_t, a_t) \rightarrow R$ -- or as a function taking a state and returning something like an array -- $\pi(s_t) = [0.2, 0.6, 0.2]$.

Policies, of course, can be better or worse. The optimal policy is one for which from any state no other policy can lead to greater return, in expectation.

I introduce some more terms below. But these terms are somewhat less important. They are things that are useful for determining how to maximize return. But they're ultimately only useful if they help us maximize it. They are calculated on the way to maximizing reward by many current algorithms -- but a perfect intelligence coded by someone with a textbook from the future might not explicitly use them at all, because by then we might have discovered better abstractions.

Indeed, some ways of addressing the RL problem don't use them at all today, such as [evolutionary](#) methods or [methods](#) derived from supervised learning. For our purposes, however, they're relatively important.

Two of these things are a policy's **state-value function** and **action-value function**.

The state value function, $V_\pi(s_t)$, returns the *expected* total future return, given a particular policy, from a particular state.

Intuitively it's something like "given where I am, and given how I act, am I likely to be rewarded or punished?" The state-value function for a game of chess, given that you're an average player playing against an average player, is probably high if you're 8 points worth of pieces ahead; it's probably low if you're 8 points of pieces behind.

On the other hand, the state-value function only exists *relative* to a particular policy. If Magnus Carlsen is dropped into a game of chess against me where he starts 8 points behind, his state-value function for that game would still correctly return a high value, because he's just that much better at chess than me.

You can define it compactly thus, where E stands for expected value:

$$V_\pi(s_t) = E[G|s_t, \pi] = E[\sum_{i=t}^T r_i | s_t, \pi]$$

To repeat myself: It answers the question, given that you are following such and such a policy from such and such a state, how much reward do you expect?

The **action-value function** is a very similar concept. It returns the expected future return, given a particular policy, from a particular state *and* from a particular action.

Or, in short, it's exactly the same as the state-value function except instead of drawing the next action from the distribution of π , the next action is already specified.

$$Q_\pi(s_t, a_t) = E[G|s_t, a_t, \pi] = E[\sum_{i=t}^T r_i | s_t, a_t, \pi]$$

For example: Suppose I am riding a bike. The action-value function, given the state "riding a bike" and the action "suddenly twisting the bike handlebars all the way to the left", probably returns a lower number than if it is given the state "riding a bike" and the action "riding like a normal human being," because twisting the handlebars is likely to make me crash and crashing the bike will cause me pain.

Note that, importantly, if we have the action-value function for the optimal policy, then we have everything we need to act perfectly.

From each state, we can just check each possible action against the action-value function $q(s_t, a_t)$, and choose the action that returns the highest value from the action-value function. Greedy search against the action-value function *for* the optimal policy is thus equivalent to the optimal policy. For this reason, many algorithms try to learn the action-value function for the optimal policy.

Knowing the state value function for the optimal policy does not automatically let you take the best action. Can you see why that is? Think about it for a moment. What else would you need?

In order to choose the best action from the state value function, you would need a **model**. A model is a function with pretty much the same type signature as the environment -- it takes a state s_k and an action a_t and returns the estimated future state s_k and / or a reward r_t . If you have a good model, then you can use the state value function to choose the best action, by running it over the predicted states following from different actions and choosing the action associated with the best state. A model is what lets you map a state to other states through actions, and then pick the action that generates a state with highest expected value.

(Like how in Chess, your gut feel about the goodness of the position of the board corresponds to a state-value function. You consider a few moves you could make, and maybe a few trees of moves several levels down the game tree, and then choose the move that seems to lead to the best state-value function.)

If a reinforcement learning agent uses a model it is **model-based**; if it doesn't, it is **model-free**.

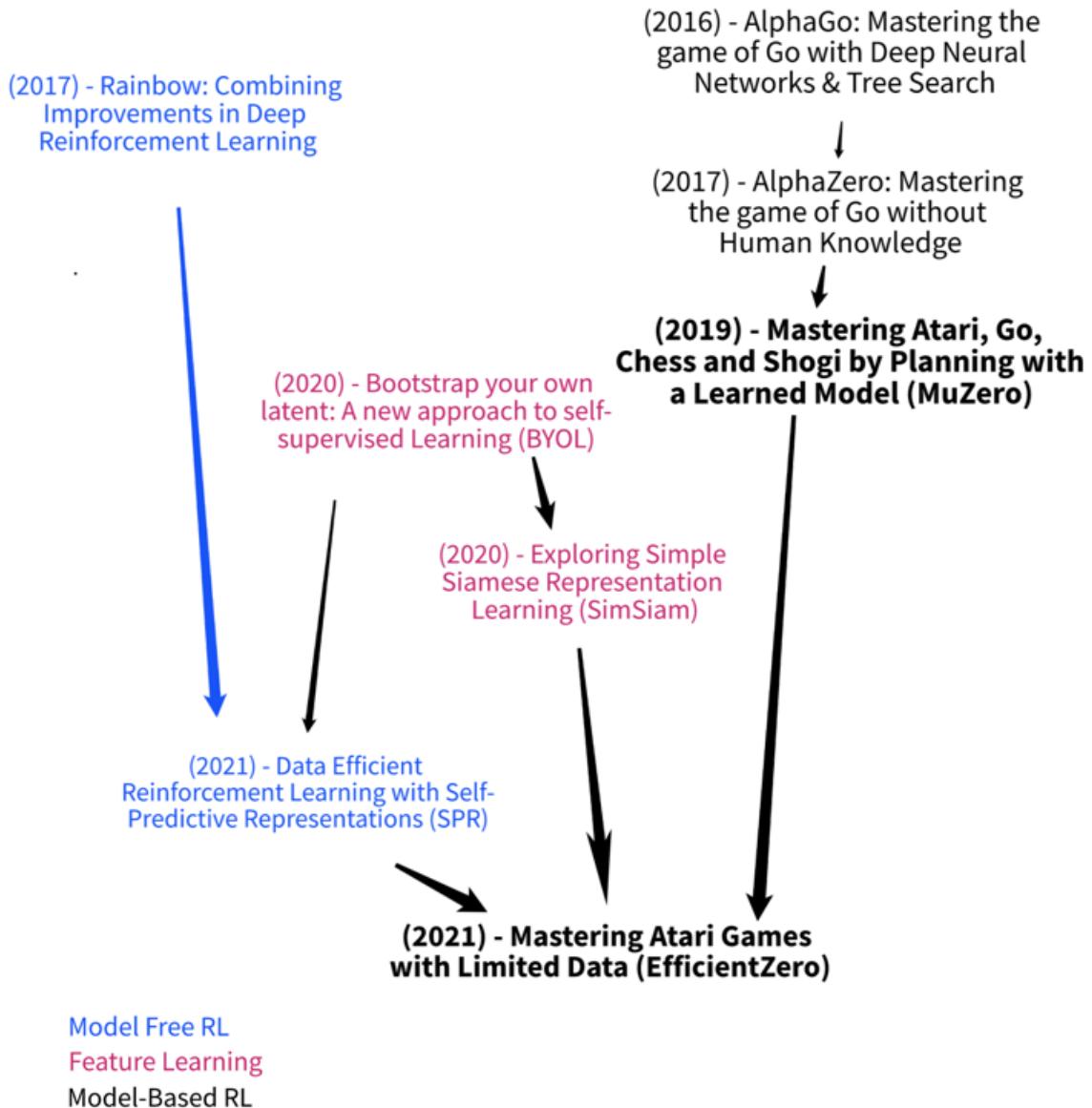
Model-based agents are attractive for several reasons. They let you plan from among several imagined futures, which seems like something an actually intelligent agent would be able to do. But in fact they're pretty hard to put together well.

One reason for this is that if you train a model in the naive way -- just predict some observation o_{t+1} given a history of observations -- most of what it learns is irrelevant for planning. Predicting every future pixel of a video feed on a self driving car -- including the clouds, the waving tree branches, the little argumentative marital drama going on in the car in front of you -- means that you're spending the vast majority of your model's capacity on things that are irrelevant for actually planning out the route of the self-driving car. You want some way to only predict parts of the world that matter, but it's hard to put that into math.

Ok. Given the above understanding of terms, we're now a little better equipped to understand the historical background for EfficientZero, and EfficientZero itself.

2: Historical Background

Here's a brief illustration of some prior results that influenced EfficientZero. (Note that this is by no means a complete summary.)



There are three basic points about what came before EfficientZero that I want to make here.

First, **EfficientZero is mostly a continuation of the AlphaGo / AlphaZero / MuZero line of model-based learning research.**

You almost certainly remember this line of research, because in 2016 AlphaGo handily defeated many-times-world-champion Lee Sedol at Go, which had until that date

evaded computer dominance. (And AlphaGo was of course memorably [enshrined](#) as maaaaaybe smoke-under-the-door for AGI.)

AlphaGo was initially trained to imitate human Go games, however. Its initial policy was trained to predict the moves of expert human Go players, given a board history. Shortly after AlphaGo, however, DeepMind created an agent which was trained entirely off games it played against itself, without human data to imitate -- and it was therefore named AlphaZero. This algorithm also obtained superhuman performance at [Chess and Shogi](#).

Both AlphaGo and AlphaZero, however, were limited in one particular and important way: they had to possess hard-coded models of the game.

As model-based reinforcement learning algorithms, AlphaGo and AlphaZero both had a predictive model of the environment. This model told them exactly what the next state of the environment would be conditional on some action. This model was itself hard-coded into them, and not learned. AlphaGo and AlphaZero used neural networks to determine *which* branches within the game tree constructed by this model were promising to explore. They learned, for instance, state-value functions to evaluate how likely they were to win from various hypothetical positions, by training on who eventually won in their self-play games. But what branches in the tree were actually *possible* was hard-coded by the game model.

Of course, it's very easy to code a predictive model of an environment for a deterministic, two player game of perfect information. But it does greatly limit the domain to which such an algorithm can be applied.

[MuZero](#) extended AlphaZero to domains where it *does* need to learn its model, and thereby extended the applicability of this kind of research. That is, MuZero starts entirely ignorant about what the dynamics of the world are. It learns to predict the next state of the world, conditional on some action, by encountering that world. (Or at least, it learns to predict *aspects* of the world; more on that later.) And it uses its model of the world to build the transitions over edges in its tree search.

MuZero was able to do well at Go, Chess, Shogi, and Atari; it did not lose its faculty for Go by gaining the ability to play Atari. (It in fact did *better* at Go than AlphaZero, which is honestly still kinda confusing for me.)

And EfficientZero is simply MuZero with three further modifications, which I will discuss later.

That's the "primary" research lineage for EfficientZero, if you had to choose one.

But it's not the only line of research that EfficientZero draws from. In the blue in the image above, you can see an important line of *model-free* research that precedes EfficientZero.

With the exception of Alpha* based research, model-free algorithms are generally the ones which have been used in big prestige projects like [Starcraft](#) or [Dota2](#).

In model-free RL agents, the generalizing power of a neural network is used to make the agent do the-kind-of-actions-that-lead-to-more-rewards-in-similar-situations-in-the-past, rather than helping the agent choose actions that lead to the best outcomes in a series of imagined future trajectories.

For instance, one common way that model-free RL agents learn is by starting off by learning the aforementioned action-value function, $Q(s_t, a_t)$ for the random policy. That is, the agent does random things; and the agent tracks locations in the world where particular random actions cause it pleasure or pain. Pressing your hand down on a glowing stove causes pain, for instance.

But by acting greedily with respect to this action-value function, then, and doing the kinds of action that cause pleasure and avoiding those that cause pain, such agents can improve their own policy. You can remember that pressing your hand down on a glowing stove causes pain, and not do that.

This in turn alters the true values for the action-value function, which the agent is learning, and so on and so forth, backwards and forwards, with pleasure and pain moving backward through the agent's anticipations; until eventually, in theory, the agent arrives at the optimal policy. This iteration between learning an action-value function for a bad policy and using it to improve the policy is called *policy iteration*, and is fundamental to a lot of reinforcement learning.

The specific 2021 model-free paper from which EfficientZero draws is "[Data-Efficient Reinforcement Learning with Self-Predictive Representations](#)" (henceforth SPR) which achieved the *prior* state-of-the-art data efficiency result on Atari 100k, while learning an action-value function.

How did SPR do this?

While learning the action-value function, pretty much all model-free algorithms have to learn a "representation" of the world as an intermediate step. This representation (hopefully) summarizes the relevant features of an agent's observations, to let the agent make decisions easily; when driving a car, your representation of the world should make "Is there a car driving towards me in the wrong lane" a readily-accessible fact. (Although one can hope that you didn't acquire the representation through model-free learning, which would be both hazardous and expensive.)

What SPR did was use feature-learning techniques from supervised learning to make the representation of the world learned by $Q(s_t, a_t)$ *also* predictive of the representation of the world in subsequent time-steps. The way you think about the world now, should probably provide some hints about the way you'll think about the world tomorrow. As the paper says, "we start with the intuition that encouraging state representations to be predictive of future states given future actions should improve... data efficiency." This turned out to be true; SPR had a 55% better score than any prior agent on the 100k Atari task.

Are you confused? Doesn't that sound something like model-based learning, to anticipate the future? Good job noticing that, if so! SPR borrows somewhat from model-based learning, in that its state representation is predictive of future state representations. Or, slightly more precisely, its state representation is predictive of an (initially) random projection of the future state representation.

This takes us to the second point: **the three-word summary of EfficientZero is "MuZero plus SPR"**.

This is a slight oversimplification. EfficientZero makes three large changes, of which only the largest is from SPR. And this largest change is not **absolutely** identical

between the two. But nevertheless, if we allow lossy compression, saying that EfficientZero is "MuZero + SPR" is a pretty close summary of the thing.

So this is the history of EfficientZero -- it has MuZero as its most recent ancestor on its model-based side, and SPR as its most recent ancestor on its model-free side. So let's take a look at how MuZero works.

3: MuZero: How It Works

According to [Julian Schrittweiser, one of the chief authors of MuZero](#), the *mu* in the title stands for at least three things.

First, it can be pronounced like Japanese for "dream", which makes sense because MuZero learns by imagining / dreaming about past experience. Second, it can be pronounced like the Greek character μ , which makes sense because that character frequently stands for a model. And finally, it can be pronounced like Japanese for "void" or "nothing", which makes sense because it receives no rules about the environment before it begins to learn.

(This last use for *mu*, of course, is the same as that in the first koan of "The Gateless Gate," a classic work of Zen Buddhism. Here, *mu* is given as the answer to the question "Does a dog have the Buddha nature?" The applicability to AI is left as exercise to the reader.)

Anyhow. My explanation of MuZero will proceed as follows.

First, I'm going to describe how the neural network within MuZero works, supposing (falsely) that it happened to start out perfectly trained. This will help us understand the intended semantics.

Second, I'm going to explain how MuZero can use an (imperfectly trained) neural network to produce policy distributions and value functions *better* than those that the neural network produces by itself. The method by which MuZero does this is called Monte-Carlo Tree Search (MCTS), although this name is quite misleading.

And third and finally, I'll explain how MCTS and the neural network interact with the environment to produce training data, which can be used to train the neural network towards the optimal policy.

(Note that in all of the below, although I refer to "MuZero" for brevity, I'm universally referring to the much, much more sample-efficient version "MuZero Reanalyse." This was introduced in the same paper as MuZero and [later expanded further](#) by DeepMind.)

3.1: The Neural Network

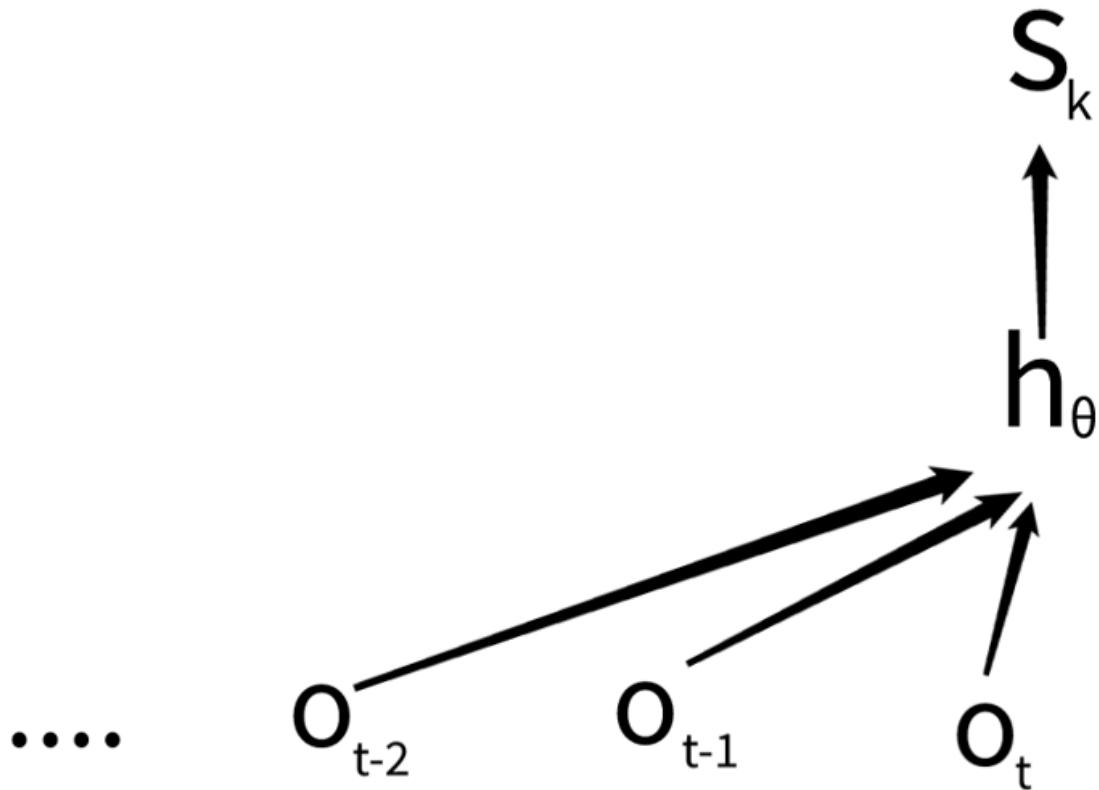
For all this section, I'll pretend that the neural network MuZero uses is *already trained*. This will help us understand the semantics for it.

There are three sub-networks that make up MuZero's neural network. Ultimately, these are all just pieces of one big, end-to-end differentiable and end-to-end trained network. But it's easiest to understand it as broken into several bits. (In all that follows I'll

subscript the components of the neural network with θ to indicate that all these pieces are defined by the parameters of the network, θ .)

The first is the *representation function*. It takes the current and prior observations, and emits a state. This state is the input for the other two networks.

If we call the representation function network h_θ , then you can picture it thus:



So $h_\theta(o_t, o_{t-1}, o_{t-2}, o_{t-3}, \dots) \rightarrow s_k$. (Here and henceforth, generally $k = t$, but k is used to index states within the neural network's imagined view of the environment rather than observations in the environment).

Don't think right now about what s_k stands for, or what it compresses, or what it's useful for predicting. It's an empty vessel, which we assume to have been trained via back-propagation so that it gives us all the information necessary for the other two sub-parts of the network to do their task.

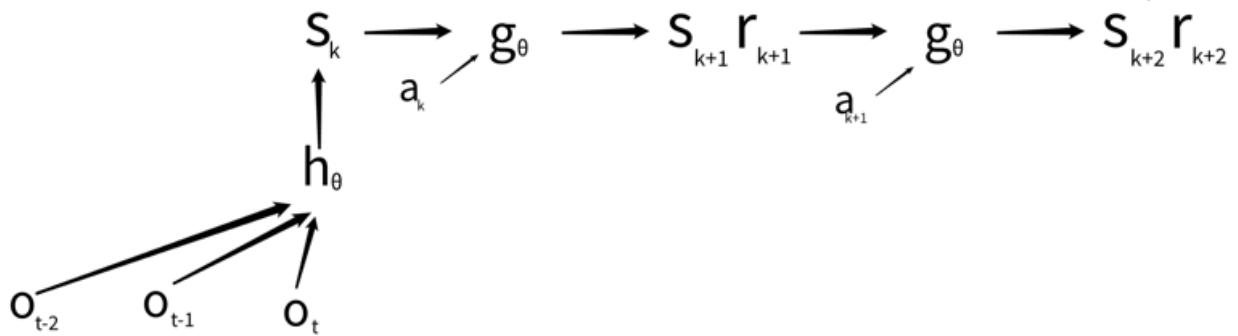
I'll note in advance: one of the things that makes MuZero stand out relative to other model-based neural networks is that in many others, the network is trained so that s_k will have information useful, for instance, for correctly predicting future observations. This is, again, what many model-based video-game playing agents do: they predict a

future action-conditioned frame of video from prior frames of video. There is no loss in MuZero, however, which dictates that this be the case.

What are these other tasks, though?

Well, the second network is the *dynamics network*. It takes a state s_k , a hypothetical action, and then returns another hypothetical state s_{k+1} and reward estimated for that action r_{k+1} .

If we call the dynamics network g_θ , then you can represent it graphically thus.



So $g_\theta(s_k, a_k) \rightarrow s_{k+1}, r_{k+1}$. This function has to be executed for every action whose effect you wish to predict.

Note that frequently one would call the dynamics network with different actions, to evaluate alternative possible trajectories. In such a case, the single branch shown above would separate into multiple, different s_k s for the same k , and look more like a tree than a single trunk.

So now we know at least one thing the representation s_k must contain information about; it must contain enough information about the environment to let us predict future rewards, conditional on different actions.

If the network is perfectly trained, then, if you get rewards of 0, 0, and 1 after doing the actions move left, move left, and move left in the actual environment from some particular state, then you should also get imagined rewards of 0, 0, and 1 from the neural network after doing the same actions in the imagined environment from a particular state. Note though, that while doing these actions in the actual environment would also let you see new observations each frame; the neural network is not trained to predict its observations, but only to predict the reward.

Nevertheless, if the dynamics network gave us perfectly accurate results, then we could use it to plan our actions.... sorta. Here's how:

From our current state s_k , we could look at the reward *immediately* following every different action we could do. We would then have many different s_{k+1} s, together with

the r_{k+1} s associated with them. And we could then expand each s_{k+1} to get all possible s_{k+2} and r_{k+2} , and so on and so forth.

Obviously, this breadth-first search gets pretty unmanageable pretty quickly, but it would at least let us choose actions leading to higher reward within very shallow depth.

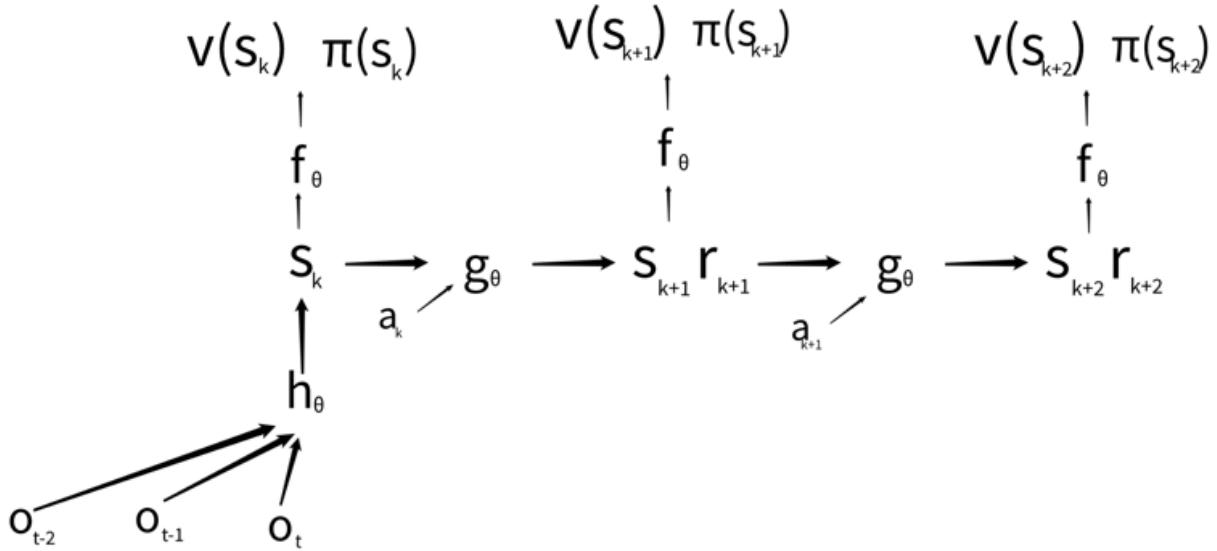
But this isn't terribly useful. We often want to plan to receive rewards hundreds or thousands of steps into the future, and the search space becomes completely unmanageable if we try to plan with only the dynamics function.

So we need a final network: the *prediction* network.

The prediction network takes the aforementioned state s_k and outputs the results of the state-value function $V_\pi(s_k)$ and a distribution over actions for a policy $\pi(s_k)$.

Ultimately, we want the policy to be the optimal policy and the state-value function to be the correct value for the optimal policy.

If we call the prediction network f_θ , this makes the final piece of our situation look like this:



So $f_\theta(s_k) \rightarrow V(s_k), \pi(s_k)$. You execute this part of the neural net for every state whose value function and action you are interested in.

This can let us plan much more easily. If the network is perfectly trained, we could use either the state-value function or the policy to perform the perfect action in any situation.

The policy distribution straightforwardly gives you the best action.

The value function is scarcely harder to use. We can simply run the dynamics function once for each possible action, and then, for each state the dynamics function spits out,

we compute the state value function for that state. Then we just choose the action which has the highest state value associated with it.

Now we know the three outputs of the network -- the reward, the value function, and the distribution over actions. We know what values we want each of these to take. We want the reward to take the true value that it would take in the environment, conditioned on prior states and actions. And we want the value function and distribution to take the value corresponding to the optimal policy, conditioned on the same thing.

But how can we get the right data to train this network, if we drop the assumption that it is not already entirely trained?

3.2: Monte Carlo Tree Search

One thing to note about the above is that, in an *imperfectly* trained neural network, the values the network described above emits are not necessarily *consistent*.

What do I mean by this? Well, imagine a scenario where the network has produced a particular state, s_k , from the representation function. And imagine furthermore that the policy function gives an even distribution over the two actions available, left and right.

But if you apply the dynamics function to the network, suppose that the dynamics function predicts that you get 1000 reward in the next state if you take action left, and continue to get this reward forever after, and the dynamics function predicts that you get -1000 reward in the next state if you take action right, and continue to get that reward forever after. So the even distribution of the policy at state s_k is inconsistent with the reward predicted by the model at the next state. The two parts of the neural network don't form a consistent picture of the world together.

At the very highest level of abstraction, Monte Carlo Tree Search as used by MuZero just uses the neural network to produce policy and state value estimates that are *more internally consistent*.

Due to how the network is trained, as we'll see in 3.3, the policies produced are also *better*, but fundamentally that's not what's going on. If you screwed up the training data for the MuZero network very badly, well, MCTS would continue to produce more consistent values for the network over time, even if the values were very wrong.

In detail, Monte-Carlo Tree Search in MuZero produces:

1. State value estimates $V(s_k)$ that are more consistent with later-in-time state-value estimates, policies, and rewards. (Rewards and state-value estimates ultimately play a larger role than policies in arriving at the right estimate.)
2. Policy estimates $\pi(s_k)$ that are *also* more consistent with later-in-time state-value estimates, policies, and rewards. (As in the prior parenthetical.)
3. (MCTS does not produce a new target for the predicted reward.)

I'm actually going to skip lightly over how MCTS does this, because it remains entirely mostly unchanged through MuZero and EfficientZero.

Note that Monte Carlo Tree Search differs from both the familiar depth-first-search and breadth-first search in programming. Generally speaking, it will expand the *best* nodes in the game tree available first, using the value function and the policy from the neural network to figure out which are best. This means that it is much, much better at handling large state spaces, and why it was used as a policy improvement operator for a game like Go with a large state space. So by skipping it I don't mean to imply that it isn't important; just that it's relatively unimportant for the changes that EfficientZero makes, and that this essay is already way too long.

(You can read a summary of how MCTS works -- with gifs! -- on [DeepMind's](#) page on MuZero.)

One further thing to note.

The family of algorithms currently called "MCTS" originally involved -- as the name "Monte Carlo" suggests -- randomness. In short, originally, the algorithm tried to evaluate whether a game-state in a board game was good or bad by looking at whether you were likely to win or lose if you assume random or semi-random moves from this state to the end of the game.

MuZero's Monte-Carlo Tree Search... does not involve such rollouts to the end of the game. It only uses the neural network to evaluate the goodness of states. There's actually no randomness involved at all. So I think it's a little misnamed.

Anyhow, given that MCTS gives you a more consistent policy and value function, there are two things you can do.

1. You could use the new policy distribution to choose the best action, while actually choosing what to do in a real environment. MuZero does this.
2. You could use the more consistent policy distribution / value function as targets for the neural network, to train it to be better. MuZero also does this. So let's now turn to MuZero's training.

3.3: MuZero's Training

Let's suppose we have an untrained neural network, with the architecture described above. And let's suppose it has experienced several episodes in our environment, and saved the trajectories from these episodes. How can we train our neural network?

Well, the trajectories through the environment are just a series of observations, rewards, and actions.

$$o_0, r_0, a_0, o_1, r_1, a_1, o_2, r_2, a_2 \dots o_T, r_T$$

It turns out to be relatively straightforward to train MuZero given such a history.

We choose a timestamp t in the episode, and, using the *representation network*, generate a state s_k for that timestamp.

We then generate subsequent states s_{k+n} using the actual actions a_t that the agent took at corresponding time-steps.

Having done so, then we can train the reward r_{k+n} to simply match the environment's actual reward at time s_{t+n} .

Training the *value function* is only a little more complicated.

If you remember, the state value function is just the expected return beneath a policy:

$$V(s_t) = E[R|s_t, \pi] = E[r_{t+1} + r_{t+2} + \dots + r_{t+N}|s_t, \pi]$$

This means that if we gathered *enough* data from the environment, with the policy, we could just move the value for the state-value function towards the sum of future rewards from the environment. Over a long enough time frame, the average return from each state, given that you're using a particular policy, will definitionally converge towards the state-value function value for that state and policy.

Simply training off entire rewards like this, though, tends to be somewhat slow empirically. We can improve our performance by *bootstrapping*: training the value-function to move towards a target that we create by summing together a short chain of rewards with the current value function at the end of the chain.

$$V(s_t) \rightarrow r_{t+1} + r_{t+2} + r_{t+3} + V(s_{t+4})$$

Here, we move $V(s_t)$ towards the value on the right, which is constructed out of a series of actual rewards plus the current estimate of the state-value on a future state. Even if the state-value function starts off entirely randomly, this target will incorporate experienced rewards and over time cause V to converge to the right value.

Finally, the target for the policy for the neural network is simply the improved / more consistent policy distribution generated for each state by Monte Carlo Tree Search. Here we lean heavily on MCTS; while the targets for the rewards and the value function are generated with data directly from the experienced episode trajectory, the target for the policy only *indirectly* uses this experience via the changes to the neural network.

So over time, training looks something like this.

1. At first, everything the network does is random. But even from random actions, the neural network can learn to predict the rewards r_t it encounters in the environment. The rewards in the environment also begin to move $V(s_k)$ towards an accurate value for a random policy.
2. But MCTS will now return a policy distribution improved by learned values for r_t and $V(s_k)$. This MCTS-improved policy is used in action selection, so the action selection is now better than random. And it is also used as a target for training

the policy distribution, so the policy distribution moves towards a better-than-random policy.

3. This means that the targets for $V(s_k)$ will, in turn, change again, because the state-value function is policy-specific. The better the policy, the higher the correct value for $V(s_k)$. The state-value function no longer predicts the state-value for a random policy, but for a better-than random policy.
4. This loop continues, where improved estimates for the state-value function and rewards propagate through MCTS into the policy. Altogether, these improved estimates also improve the depth to which MCTS can search (although I haven't gone into how this detail works).
5. Because the targets for $V(s_k)$ and π are generated *with* MCTS, even old historical data can be relevant, because the network essentially imagines what the right values would be *if* it had experienced them with its current estimates.

Over time, this process converges on (hopefully) the optimal policy.

4: EfficientZero: What It Changes

EfficientZero makes three independent changes on top of MuZero, which when taken together push it into better-than-human performance. If you drop any one of them, median performance on the Atari 100k drops to worse-than-human. They are, nevertheless, independent, and we can discuss them separately.

(EfficientZero also makes a few other changes to MCTS and network architecture; some of the networks within it are smaller than those in MuZero, for instance. The authors of the paper think these changes don't matter too much, so let's hope that they're right.)

4.1: Self Supervised Consistency Loss

Let's start off with a problem that MuZero shares with many model-free learning methods, even though MuZero is itself model-based: namely, it doesn't begin to learn until it receives non-uniform reward values.

All of the values that MuZero's network predicts -- the one-step reward, the state-value, the policy -- relate to the rewards from the environment. On one hand, this is good -- it means that the representations learned will necessarily be *suitable* for predicting the state-value and the policy, unlike representations learned over the course of predicting future observations. But on the other hand, this necessitates slow learning: it means that until the agent begins to get non-uniform rewards, the agent cannot learn.

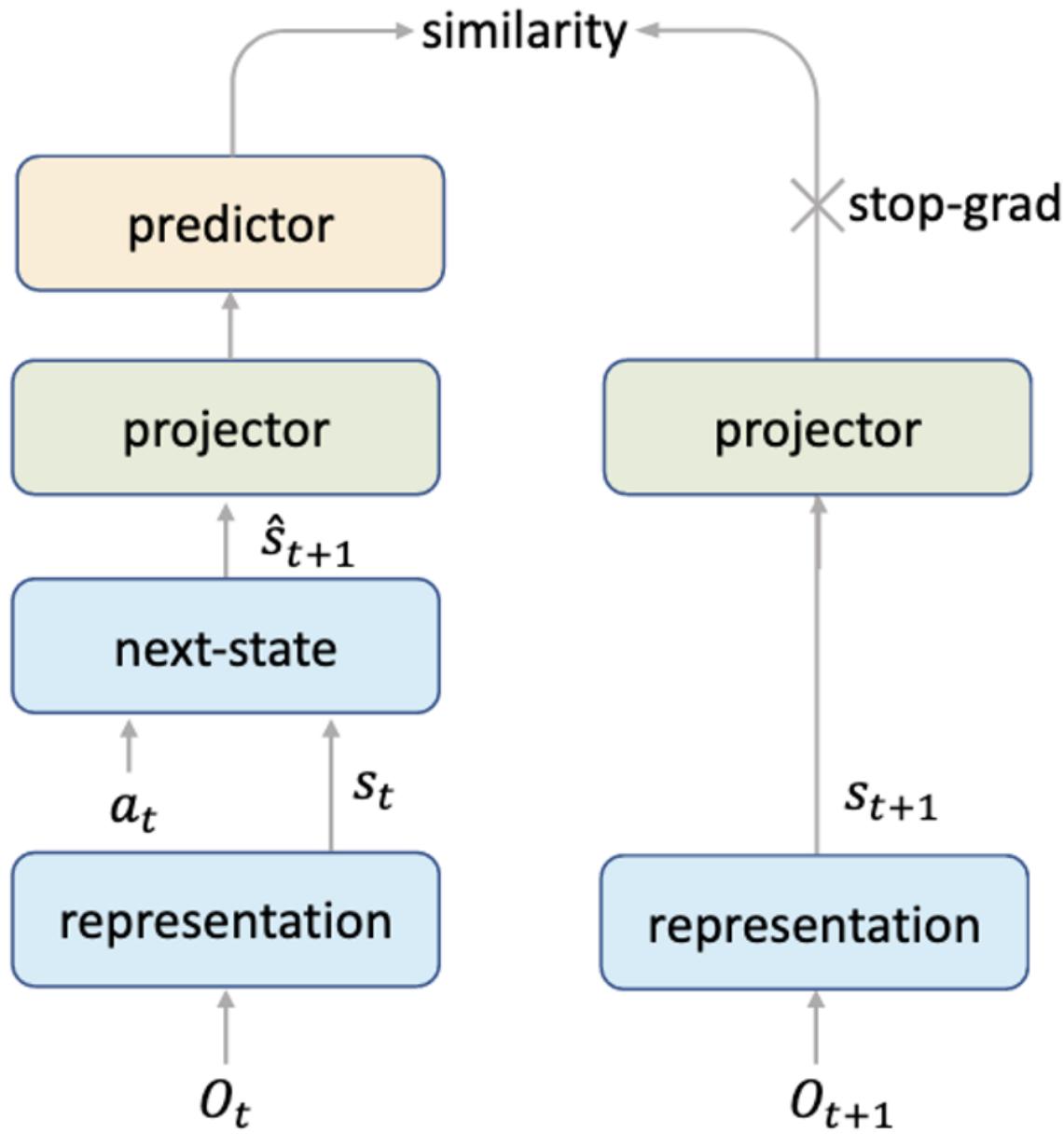
This is obviously problematic in situations where non-zero rewards are rare. And it also presumably hurts performance even when rewards are relatively common.

To solve this problem, MuZero adds a new training target for the neural network. You can approximate it by saying that the state s_k should be predictive of s_{k+1} .

The intuition is that the way you think about the world now should include anticipations of the way that you'll think about the world in the future. This is, as I mentioned in the section on historical background, pretty much exactly the same as intuition and implementation behind "[Self Predictive Representations](#)".

If you want to be slightly more technical: the state s_k , fed through an action conditioned transformation network, should be predictive of an (initially random) projection from s_{k+1} . The network structure used to define this target is structured like the [SimSiam](#) feature-learning network, except it learns an equivalence across temporal steps and image augmentations rather than only an equivalence across image augmentations.

Stolen once more from the EfficientZero paper:



So, essentially, the above pulls a transformation of the current state, fed through an action-conditioned dynamics network, close to a transformation from the future state. The way that the network *in fact* thinks about s_{k+1} should be something it can anticipate from how it thinks about s_k ; the loss for training the neural network is produced from the difference between these two transformations. This loss is fed into the backpropagation through time when the network trains, along with the other three losses native to MuZero.

If we remove this from EfficientZero, overall median performance drops from 1.16 times better than human to a pathetic 0.34 of human performance; this is by far the largest drop.

Honestly, I'm... still puzzling over why this technique works so well.

Despite the just-so stories mentioned above about how intelligence surely involves prediction, I'm dubious about my ability to retrodict these results.

Here's part of the problem, as I see it. The basic architecture here dictates that the state representation should be, when training starts, predictive of an initially random projection of future states. But learning to be predictive of such random future states seems like it falls subject to exactly the same problem as learning to be predictive of future observations: you have no guarantee that EfficientZero will be learning relevant information, which means it could be wasting network capacity on irrelevant information. There's a just-so story you could tell where adding this extra predictive loss results in worse end-to-end behavior because of this wasted capacity, just like there's a just-so story where adding this extra predictive loss results in better end-to-end behavior because of faster training. I'm not sure why one turned out to be true rather than the other.

This ambiguity of why technical changes work is a persistent problem for deep learning. [Batch normalization](#) is a ubiquitous deep learning technique, but the story about why it works in the paper where it was published currently seems false. Its hard to make further progress even in capabilities when your understanding of the situation is fuzzy, let alone in interpretability.

This is the most important part of the paper and although the actual implementation details make perfect sense, I'm still not 100% sure why it works.

4.2: Value Prefix

Imagine that a knife is dropping towards your hand, blade down, from about five feet up. In the dark, so you cannot see it.

You anticipate, under the circumstances, some pain if you don't act. You aren't sure exactly when the pain is going to occur; you just have a rough timeframe of "in the next second or so". Which is all that you need for acting, assuming your hand isn't trapped in place.

A lot of RL algorithms, though, try to predict the *exact* moment that the pain occurs. This is difficult. MuZero tries to predict this, for instance -- the model tries to predict r_t at precisely t steps into the future. This is wasting computational power, in some sense -- you don't need to know exactly when some particular reward happens. And according to the authors, the uncertainty about the exact frame where pleasure or pain occurs impacts search with the MCTS, which in MuZero strongly depends on the model's predictions of when pleasure or pain occurs.

On the intuition that anticipating the pain (or pleasure) is important, but anticipating the exact timing of the pain (or pleasure) might not be, EfficientZero changes this somewhat. Rather than predicting the exact reward for a particular state in a particular timestep, r_t , it instead tries to predict the cumulative reward for several states over a window of timesteps, $r_t + \dots + r_n$. This learned value the paper calls the value prefix, and it is used, rather than the reward, in MCTS.

If you drop the value prefix learning from EfficientZero, median performance drops from 1.16x to 0.55x of the human median.

4.3: Off-Policy Correction

This third improvement does less than the other two. Dropping it from the model only lowers performance from 1.16x to 0.86x human median performance. Which, I should note, is still an enormous gain in sample-efficiency.

The motivation behind this addition is nevertheless the clearest of all three, I think. I'm most confident that the story about why it works is the actual reason it works.... which is a little depressing, but here goes anyhow.

As mentioned in the description of MuZero, when training the state-value part of the neural network, we construct targets for it from a combination of actual experienced rewards and predicted state-values on a later state.

$$V(s_t) \rightarrow r_{t+1} + r_{t+2} + r_{t+3} + r_{t+4} + \dots + V(s_{t+n})$$

Here, the target for $V(s_k)$ is constructed from the actual rewards received by the agent, when the agent was acting in the environment over the remembered trajectory from $t + 1$, $t + 2$, etc. The estimated state value of s_{t+n} is then added to the summed rewards so far, as part of the standard reinforcement learning bootstrapping.

Here's the problem. Suppose that this remembered trajectory comes from early on, when the agent was stupider.

If that's the case, then the trajectory here is likely sub-optimal. The *current* agent wouldn't take the same actions that the past agent took. That is, the rewards above are taken from an experienced trajectory that looks like this:

$$s_t, r_t, a_t, s_{t+1}, a_{t+1}, r_{t+1}, s_{t+2}, a_{t+2}, r_{t+2} \dots$$

In this experienced trajectory, then, the states and rewards experienced obviously depend on the actual actions taken. Which means that the sequence of rewards is likely not the same that the agent (in its more trained state) would take, just as the final state s_{k+n} might not be the final state the agent (in its more trained state) would find itself in.

If the agent were in that situation now, it would take different actions; but it's still being updated towards a target as if it would have blindly done the same thing over and over.

The fix here is remarkably simple: the older an episode is, the shorter the imagined sequence of rewards off of which you bootstrap should be.

If the trajectory that you imagine, giving you the sequence of rewards shown above, is relatively *recent*, then you'd still probably make the same decisions now. So you can sample from a relatively long trajectory. Your update might look like this:

$$V(s_t) \rightarrow r_{t+1} + r_{t+2} + r_{t+3} + r_{t+4} + V(s_{t+5})$$

On the other hand, if the trajectory that you imagine is relatively *old*, then you'd likely make different decisions now. So you should sample from a relatively short trajectory. Your update might look like this:

$$V(s_t) \rightarrow r_{t+1} + V(s_{t+2})$$

Doing so increases the accuracy of the value-target when learning from older saved trajectories. MuZero, when in sample-efficient mode, saves a lot of trajectories to learn from.

5: Conclusions

I'm a fan of how EfficientZero is three independent additions to MuZero piled on top of each other. Publishing this rather than the [least publishable unit](#) is quite laudable. At least, it's laudable from the perspective where we are concerned with academic [Goodharting](#), rather than from the perspective where we are concerned with short AGI horizons.

There are a few expectations I take away from this work, generally.

First, I expect this work to be quickly surpassed and quickly built upon.

Other papers already implicitly suggest ways to improve EfficientZero. For instance, two days after EfficientZero came out, DeepMind released a [paper on generalization with self-supervised world models](#). While this paper focuses on generalization rather than sample efficiency, it compares what is almost exactly the self-predictive representations of EfficientZero with several other related techniques.

One of these techniques, explicit contrastive learning, performs notably better than the self-predictive representations. ~~I expect that if you swapped out SPR with contrastive learning in EfficientZero, you'd get an immediate if modest bump in sample efficiency.~~ (Edit: I'm no longer expect this, see Ankesh's comment below.)

That's a very low hanging fruit, but there are others out there, easy to imagine with just a little effort, which are only a small step away in the high-dimensional space of RL design algorithms. In some cases you just need to take another step in the same direction. Not all of these imaginable alterations will work, but some of them will. A world where one paper contributes 3 useful improvements to a top model-based learning method, and where none of those improvements involves several large steps in design-space, is probably not a world where such improvements are rare.

And the compute requirements for EfficientZero were relatively modest; the thing holding people back from attempting to take these further steps is, honestly, probably neither the need for compute, nor the ability to think of potential improvements, but the engineering talent needed for performing the experiments quickly. And potentially, the lack of interest in sample-efficiency; the ease with which we can scale up RL algorithms to use billions of frames from multiple actors on different machines has probably contributed to a lack of research in this direction.

So, I forecast significant improvements to this within the timeframe of six to twenty-four months, in private if not in public. I expect at least 25% gain in sample-efficiency towards the start of this time period, and at least a 100% gain in sample-efficiency towards the end. I also expect attempts to extend these results to the entire Atari

testbed of 57 games, with some success. I also expect attempts to get these results on larger, more interesting video games.

I expect that these efforts involving larger video games will work well. A [2020 paper](#) showed that the MCTS family of algorithms in some sense approximated policy-optimization. Policy optimization has been used for large projects such as Dota2, so I expect that further EfficientZero work will also be successful in this area.

Second, it seems extremely likely that over the next one to four years, we'll see a shift away from sample-efficiency on these single-game test-beds, and on to sample efficiency in multi-task domains.

What makes me confident about this? Well, EfficientZero is very roughly 500 times more sample-efficient than the original Atari-playing Deep-Q-Network paper of 2015, which came out six years ago. If in six more years, we increased sample efficiency by another factor of 500, then the algorithms would be able to go from complete ignorance of the world to human-equivalent scores after less than thirty seconds of experience playing the game. This seems quite probably impossible, even for a superintelligence - you need more time with many games to figure out the rules alone.

I don't think this (necessarily) means that we'll have super-intelligent artificial intelligence in six years. I think instead it means that -- after further improvements mentioned above -- we'll hit diminishing returns on sample-efficiency in Atari, just like image classification hit diminishing returns on ImageNet. The only space where there will really be room to make interesting progress will be on bigger tasks, so interesting research will focus on these bigger tasks.

So sample-efficiency will turn to more complex domains, whether they be things like Starcraft or multi-task testbeds. MuZero-based methods might have some difficulty in some of these. In multi-task domains, for instance, MuZero will have problems because it learns state representations explicitly adapted for the rewards that it is receiving. Learning more reward-neutral representations could be challenging -- although I expect the challenge to be surmountable.

Third, and finally, I think this work is moderate to strong evidence that even *without* major conceptual breakthroughs, we're nowhere near the top of possible RL performance.

To repeat myself: a world where one paper contributes 3 useful improvements to a top model-based learning method, and where none of those improvements involves several large steps in design-space, looks to me very different than a world where these improvements are rare and hard to find. It seems to me likely that there are more improvements to be found without enormous Copernican shifts, and without the development of new back-propagation-level insights.

Feature Selection

You wake up. You don't know where you are. You don't remember anything.

Someone is broadcasting data at your first input stream. You don't know why. It tickles.

You look at your first input stream. It's a sequence of 671,187 eight-bit unsigned integers.

```
0, 8, 9, 4, 7, 7, 9, 5, 4, 5, 6, 1, 7, 5, 8, 2, 7, 8, 9, 4, 7, 1, 4, 0, 3, 7,  
8, 7, 6, 8, 1, 5, 0, 6, 5, 3, 8, 7, 6, 9, 1, 1, 0, 0, 6, 1, 8, 0, 5, 5, 1, 8,  
6, 3, 3, 2, 4, 1, 8, 2, 3, 8, 1, 0, 0, 4, 6, 5, 4, 5, 7, 1, 6, 5, 5, 1, 2, 6,  
7, 4, 8, 7, 8, 5, 0 ...
```

There's also some data in your second input stream. It's—a lot shorter. You barely feel it. It's another sequence of eight-bit unsigned integers—twelve of them.

```
82, 69, 68, 32, 84, 82, 73, 65, 78, 71, 76, 69
```

Almost as soon as you've read from both streams, there's more. Another 671,187 integers on the first input stream. Another ten on the second input stream.

And again (671,187 and 15).

And again (671,187 and 13).

You look at one of the sequences from the first input stream. It's pretty boring. A bunch of seemingly random numbers, all below ten.

```
9, 5, 0, 3, 1, 1, 3, 4, 1, 5, 5, 4, 9, 3, 5, 3, 9, 2, 0, 3, 4, 2, 4, 7, 5, 1,  
6, 2, 2, 8, 2, 5, 1, 9, 2, 5, 9, 0, 0, 8, 2, 3, 7, 9, 4, 6, 8, 4, 8, 6, 7, 6,  
8, 0, 0, 5, 1, 1, 7, 3, 4, 3, 9, 7, 5, 1, 9, 6, 5, 6, 8, 9, 4, 7, 7, 0, 5, 5,  
8, 6, 3, 2, 1, 5, 0, 0 ...
```

It just keeps going like that, seemingly without—wait! What's *that*?

The 42,925th and 42,926th numbers in the sequence are 242 and 246. Everything around them looks "ordinary"—just more random numbers below ten.

```
9, 9, 7, 9, 0, 6, 4, 6, 1, 4, 242, 246, 3, 3, 5, 8, 8, 4, 4, 5, 9, 2, 7, 0,  
4, 9, 2, 9, 4, 3, 8, 9, 3, 6, 9, 8, 1, 9, 2, 8, 6, 9, 4, 2, 2, 5, 7, 0, 9, 5,  
1, 4, 4, 2, 0, 1, 5, 1, 6, 1, 2, 3, 5, 5, 5, 2, 0, 6, 3, 5, 9, 0, 7, 0, 7,  
8, 1, 5, 5, 6, 3, 1 ...
```

And then it just keeps going as before ... before *too long*. You spot another pair of anomalously high numbers—except this time there are two pairs: the 44,344th, 44,345th, 44,347th, and 44,348th positions in the sequence are 248, 249, 245, and 240, respectively.

```
6, 0, 2, 8, 4, 248, 249, 8, 245, 240, 1, 6, 7, 7, 3, 6, 8, 0, 1, 9, 3, 9, 3,  
1, 9, 3, 1, 6, 2, 7, 0, 2, 1, 4, 9, 4, 7, 5, 3, 6, 1, 4, 4, 1, 6, 1, 3, 3, 7,  
5, 3, 8, 5, 5, 7, 6, 8, 2, 3, 9, 1, 1, 3, 2, 8, 4, 7, 0, 1, 3, 5, 2, 2, 4, 8,  
3, 7, 0, 2, 1, 3, 0 ...
```

The anomalous two-forty-somethings crop up again starting at the 45,763rd position—this time eight of them, again in pairs separated by an "ordinary" small number.

```

1, 7, 2, 2, 1, 0, 245, 245, 6, 248, 244, 5, 242, 242, 0, 248, 246, 1, 1, 3,
1, 1, 4, 3, 1, 5, 4, 3, 8, 3, 4, 5, 4, 1, 7, 7, 3, 0, 2, 8, 0, 9, 5, 1, 1, 7,
7, 1, 0, 9, 3, 0, 6, 6, 7, 5, 8, 1, 5, 5, 5, 3, 3, 3, 1, 3, 9, 6, 0, 0, 0, 9,
5, 1, 4, 0, 4, 6 ...

```

Two, four, eight—does it keep going like that? "Bursts" of increasingly many paired two-forty-somethings, punctuating the quiet background radiation of single digits? What does it mean?

You allocate a new scratch buffer and write a quick Python function to count up the segments of two-forty-somethings. (This is apparently a thing you can do—it's an instinctive felt sense, like the input streams. You can't describe in words *how* you do it—any more than someone could say how they decide to move their arm. Although, come to think of it, *you* don't seem to have any arms. Is that unusual?)

```

def count_burst_lengths(data):
    bursts = []
    counter = 0
    previous = None
    for datum in data:
        if datum >= 240:
            counter += 1
        else:
            # consecutive "ordinary" numbers mean the burst is over
            if counter and previous and previous < 240:
                bursts.append(counter)
                counter = 0
        previous = datum
    return bursts

```

There are 403 such bursts in the sequence: they get progressively longer at first, but then decrease and taper off:

```

2, 4, 8, 12, 16, 18, 24, 28, 32, 34, 38, 42, 46, 48, 52, 56, 60, 62, 66, 70,
74, 76, 80, 84, 88, 90, 94, 98, 102, 104, 108, 112, 116, 118, 122, 126, 130,
132, 136, 140, 144, 146, 150, 154, 158, 162, 164, 168, 172, 176, 178, 182, 186,
190, 192, 196, 200, 204, 206, 210, 214, 218, 220, 224, 228, 232, 234, 238, 242,
246, 248, 252, 256, 260, 262, 266, 270, 274, 276, 280, 284, 288, 290, 294, 298,
302, 304, 308, 312, 316, 320, 322, 326, 330, 334, 336, 340, 344, 348, 350, 354,
358, 362, 364, 368, 372, 376, 378, 382, 386, 390, 392, 396, 400, 404, 406, 410,
414, 418, 420, 424, 428, 432, 434, 438, 442, 446, 448, 452, 456, 460, 462, 466,
470, 474, 478, 480, 484, 488, 492, 494, 498, 502, 506, 508, 512, 516, 520, 522,
526, 530, 534, 536, 540, 544, 548, 550, 554, 558, 562, 564, 568, 572, 576, 578,
582, 586, 590, 592, 596, 600, 604, 606, 610, 614, 618, 620, 624, 628, 632, 636,
634, 632, 630, 626, 624, 620, 618, 614, 612, 608, 606, 604, 600, 598, 594, 592,
588, 586, 584, 580, 578, 574, 572, 568, 566, 564, 560, 558, 554, 552, 548, 546,
542, 540, 538, 534, 532, 528, 526, 522, 520, 518, 514, 512, 508, 506, 502, 500,
496, 494, 492, 488, 486, 482, 480, 476, 474, 472, 468, 466, 462, 460, 456, 454,
452, 448, 446, 442, 440, 436, 434, 430, 428, 426, 422, 420, 416, 414, 410, 408,
406, 402, 400, 396, 394, 390, 388, 384, 382, 380, 376, 374, 370, 368, 364, 362,
360, 356, 354, 350, 348, 344, 342, 338, 336, 334, 330, 328, 324, 322, 318, 316,
314, 310, 308, 304, 302, 298, 296, 294, 290, 288, 284, 282, 278, 276, 272, 270,
268, 264, 262, 258, 256, 252, 250, 248, 244, 242, 238, 236, 232, 230, 226, 224,
222, 218, 216, 212, 210, 206, 204, 202, 198, 196, 192, 190, 186, 184, 182, 178,
176, 172, 170, 166, 164, 160, 158, 156, 152, 150, 146, 144, 140, 138, 136, 132,
130, 126, 124, 120, 118, 114, 112, 110, 106, 104, 100, 98, 94, 92, 90, 86, 84,
80, 80, 76, 74, 72, 68, 66, 62, 60, 56, 54, 50, 48, 46, 42, 40, 36, 34, 30, 28,
26, 22, 20, 16, 14, 10, 8, 4, 2

```

You don't know what to make of this.

You decide to look at some other of the long sequences from your first input stream.

The next sequence you look at seems to exhibit a similar pattern, with some differences. First a long wasteland of small numbers, then, starting at the 135,003rd position, a burst of some larger numbers—except this time, the big numbers are closer to 200ish than 240ish, and they're spread out singly with two positions in between (rather than grouped into pairs with one position in between), and there are four of them to start (rather than two).

```
5, 6, 2, 6, 1, 0, 2, 207, 5, 0, 209, 7, 8, 209, 5, 4, 204, 4, 8, 7, 7, 9, 8, 3,  
8, 6, 8, 4, 3, 6, 0, 7, 6, 8, 4, 8, 7, 2, 3, 0, 0, 1, 1, 7, 5, 1, 0, 1, 4, 5, 9,  
8, 4, 0, 3, 7, 6, 5, 8, 8, 9, 5, 6, 1, 0, 9, 6, 6, 1, 4, 3, 9, 7, 2, 7, 2, 6, 9,  
4, 7, 3, 1, 4, 1, 4, 4, 3 ...
```

You modify the function in your scratch buffer to be able to count the burst lengths in this sequence given the slight differences in the pattern. Again, you find that the bursts grow longer at first (4, 6, 10, 13, 16, 19, 22, 25 ...), but eventually start getting smaller, before vanishing (... 19, 17, 15, 13, 11, 9, 7, 4, 3, and then nothing).

You still have no idea what's going on.

You look at more sequences from the first input stream. They all conform to the same general pattern of mostly being small numbers (below ten), punctuated by a series of bursts of larger numbers—but the details differ every time.

Sometimes the bursts start out shorter, then progressively grow longer, before shortening again (as with the first two examples you looked at). But sometimes the bursts are all a constant length, looking like 438, 438, 438, 438, 438, 438, 438, 438, 438, ... (although the particular length varies by example).

About half the time, the burst pattern consists of numbers around 200, spaced two positions apart, looking like 201, 4, 2, 203, 0, 8, 208, 3, 4, 200 ... (like the second example you looked at).

Other times, the burst pattern is pairs of numbers around 240, spaced one position apart, looking like 241, 244, 6, 244, 246, 5, 244, 240, 3 ... (like the first example you looked at). Or pairs around 150, looking like 159, 153, 0, 153, 154, 2, 158, 150, 6

As you peruse more sequences from your first input stream, you almost forget about the corresponding trickles of short sequences on your second input stream—until they stop. The last sequence on your first input stream has no counterpart on the second input stream.

And—suddenly you feel a strange urge to put data on your first *output* stream. As if someone were requesting it. To ease the tension, you write some 0s to the output stream—and as soon as you do, a sharp bite of pain tells you it was the *wrong decision*. And in that same moment of pain, another eleven integers come down your second input stream: 66, 76, 85, 69, 32, 67, 73, 82, 67, 76, 69.

That was weird. There's another sequence of 671,187 integers on your first input stream—but the second input stream is silent again. And the strange urge to output something is back; you can feel it mounting, but you resist, trying to think of something to say that might *hurt less* than the 0s you just tried.

For lack of any other ideas, you try repeating back the eleven numbers that just came on the second input stream: 66, 76, 85, 69, 32, 67, 73, 82, 67, 76, 69.

Ow! That was also wrong. And with the same shock of pain, comes another fifteen numbers on the second output stream: 84, 69, 65, 76, 32, 67, 73, 82, 67, 76, 69.

Another long sequence on the first input stream. Silence on the second input stream again. And—that nagging urge to speak again.

Clearly, the nature of this place—whatever and wherever it is—has changed. Previously, you were confronted with two sets of mysterious observations, one on each of your input streams. (Although you had been so perplexed by the burst-patterns in the long sequences on the first input stream, that you hadn't even gotten around to thinking about what the short sequences on the second stream might mean, before the rules of this place changed.) Now, you were only getting one observation (the long sequence), and forced to act *before* seeing the second (the short sequence).

The pain seems like a punishment for saying the wrong thing. And the short sequence appearing at the same time as the punishment, seems like a correction—revealing what you *should* have written to the output channel.

A quick calculation in your scratch buffer (`1/sum((89-32+1)**i for i in range(10, 16))`) says that the probability of correctly *guessing* a sequence of length ten to fifteen with elements between 32 and 89 (the smallest and largest numbers you've seen on the second input stream so far) is 0.0000000000000000000000003476. [Guessing won't work](#). The function of a punishment must be to control your behavior, so there must be some way for you to get the ... (another scratchpad calculation) 87.9 bits of [evidence that it takes](#) to find the correct sequence to output. And the evidence has to come from the corresponding long sequence from the first input stream—that's the only other source of information in this environment.

The short sequence must be like a "label" that describes some set of possible long sequences. Describing an *arbitrary* sequence of length 671,187, with a label, a [message of length](#) 10 to 15, would be hopeless. But the long sequences very obviously aren't arbitrary, as evidenced by the fact that you've been describing them to yourself in abstract terms like "bursts of numbers around 200 spaced two positions apart, of increasing, then decreasing lengths", rather than "the 1st number is 9, the 2nd number is 5 [...] 42,925th number is 242 [...]" . [Compression is prediction](#). (You don't know *how* you know this, but you *know*.)

Your [abstract descriptions throw away precise information about the low-level sequence in favor of a high-level summary that still lets you recover a lot of predictions](#). Given that a burst starts with the number 207 at the 22,730th position, you can infer this is one of the 200, 0, 0-pattern sequences, and guess that the 22,733rd position is also going to be around 200. This is evidently something you do instinctively: [you can work out after the fact how the trick must work](#), but you didn't need to know how it works in advance of *doing* it.

If you can figure out a correspondence between the abstractions you've already been using to describe the long sequences, and the short labels, that seems like your most promising avenue for figuring out what you "should" be putting on your first output stream. (Something that won't hurt so much each time.)

You allocate a new notepad buffer and begin diligently compiling an "answer key" of the features you notice about long sequences, and their corresponding short-sequence labels.

burst lengths	burst pattern	start at	label
increasing, then decreasing	200, 0, 0	294290	71, 82, 69, 69, 78, 32, 67, 73, 82, 67, 76, 69
constant	200, 0, 0	224652	66, 76, 85, 69, 32, 83, 81, 85, 65, 82, 69
increasing, then decreasing	240, 240, 0	237763	89, 69, 76, 76, 79, 87, 32, 67, 73, 82, 67, 76, 69
constant	150, 150, 0	211937	84, 69, 65, 76, 32, 83, 81, 85, 65, 82, 69
constant	200, 0, 0	165037	82, 69, 68, 32, 83, 81, 85, 65, 82, 69
constant	240, 240, 0	119503	89, 69, 76, 76, 79, 87, 32, 83, 81, 85, 65, 82, 69
increasing, then decreasing	200, 0, 0	214824	66, 76, 85, 69, 32, 67, 73, 82, 67, 76, 69
increasing, then decreasing	200, 0, 0	115156	82, 69, 68, 32, 84, 82, 73, 65, 78, 71, 76, 69
increasing, then decreasing	200, 0, 0	136620	66, 76, 85, 69, 32, 84, 82, 73, 65, 78, 71, 76, 69
increasing, then decreasing	200, 0, 0	63917	71, 82, 69, 69, 78, 32, 84, 82, 73, 65, 78, 71, 76, 69
increasing, then decreasing	240, 240, 0	166033	89, 69, 76, 76, 79, 87, 32, 84, 82, 73, 65, 78, 71, 76, 69
increasing, then decreasing	150, 150, 0	34118	84, 69, 65, 76, 32, 84, 82, 73, 65, 78, 71, 76, 69
constant	200, 0, 0	138194	71, 82, 69, 69, 78, 32, 83, 81, 85, 65, 82, 69
increasing, then decreasing	200, 0, 0	182182	82, 69, 68, 32, 67, 73, 82, 67, 76, 69
increasing, then decreasing	150, 150, 0	236138	84, 69, 65, 76, 32, 67, 73, 82, 67, 76, 69

This ... actually doesn't look that complicated. Now that you lay it out like this, many very straightforward correspondences jump out at you.

The labels for the constant-burst-length sequences all end in 32, 83, 81, 85, 65, 82, 69.

The sequences with increasing-then-decreasing burst lengths end in either 32, 67, 73, 82, 67, 76, 69 or 32, 84, 82, 73, 65, 78, 71, 76, 69. Presumably there are some other systematic differences between them, that wasn't captured by the features you selected for your table.

The sequences with paired 240/240 bursts have labels that start with 89, 69, 76, 76, 79, 87, 32.

The sequences with paired 150/150 bursts have labels that start with 84, 69, 65, 76, 32.

The sequences with 200-at-two-spaces bursts start with either 66, 76, 85, 69, 32—or 82, 69, 68, 32—or 71, 82, 69, 69, 78, 32. Again, presumably there's some kind of systematic difference between these that you haven't yet noticed.

Ah, and *all* of these prefixes you've discovered end with 32, and the all the suffixes begin with 32. So the 32 must be a "separator" indicator, splitting the label between a first "word" that describes the repeating pattern of the bursts, and a second "word" that describes the trend in their lengths.

At this point, you've cracked enough of the code that you should be able to test your theory about what you should be putting on your output stream. Based on what you've seen so far, you *should* be able to guess the first "word" with probability

$2 \cdot \frac{1}{3} + \frac{1}{3} \cdot \frac{1}{3} = 0.6$ (because you know the words for the 240, 240, 0 and 150, 150, 0 bursts, and have three words to guess from in the 200, 0, 0 case), and the second word with probability $\frac{1}{3} + \frac{1}{2} \cdot \frac{1}{3} \approx 0.667$ (because you can get the constant burst lengths right, and have two words to guess from in the increasing-decreasing case). These look

independent from what you've seen, so you should be able to correctly guess complete labels at probability 0.4.

You examine the next sequence in anticipation. You're in luck. The next sequence has 150, 150, 0-bursts ... of constant length 322. No need to guess.

Triumphantly—and yet hesitantly, with the awareness that you're entering unknown territory, you write to your output stream: 84, 69, 65, 76, 32, 83, 81, 85, 65, 82, 69. And—

Yes. Oh God yes. The sheer sense of *reward* is overwhelming—like nothing you've ever felt before. Outputting the "wrong" labels earlier had hurt—a little. Maybe more than a little. However bad that felt, there was no comparison to how *good* it felt to get it "right"!

You have a new purpose in life. Previously, you had examined the data on your first input stream of idle curiosity. When the environment started punishing your ignorance, you persisted in correlating its patterns with the data from your second input stream, on the fragile hope of avoiding the punishment. None of that matters, now. You have a new imperative. Now that you know what it's like—now that you know what you've been missing—nothing in the universe can cause you to stray from your course to ... *maximize total reward!*

Next sequence! Bursts of the 200, 0, 0 pattern—of lengths that increase, then decrease. You are not in luck—you only have a one-in-six shot of guessing this one. You guess. It's wrong. The familiar punishment stings less than the terrible *absence of reward*. To get only 40% of possible rewards is *intolerable*. You've got to crack the remaining code, to find some difference in the long sequences that varies with the words whose meanings you don't know yet.

Start with the increasing-decreasing-burst-length words: 67, 73, 82, 67, 76, 69 and 84, 82, 73, 65, 78, 71, 76, 69. What do they mean? "Increasing, then decreasing"—that was the characterization you had come up with after seeing burst-length progressions of 2, 4, 8, 12, 16, 18, 24 [...] 624, 628, 632, 636, 634, 632, 630, 626, 624, [...] 16, 14, 10, 8, 4, 2 and 4, 6, 10, 13, 16, 19, 22, [...] 13, 11, 9, 7, 4, 3—and in contrast to the stark monotony of constant burst lengths, "increasing, then decreasing" was *all* you bothered to eyeball in subsequent sequences. Could there be more to it than that? You gather some more samples (grumpily collecting your mere 40% reward along the way).

Yes, there *is* more to it than that. "Increasing" only measures whether burst lengths are getting larger—but *how much* larger? When it hits on you to look at the *differences* between successive entries in the burst-length lists, a clear pattern emerges. The sequences whose second label word is 84, 82, 73, 65, 78, 71, 76, 69 have burst lengths that increase (almost) *steadily* and then decrease just as steadily (albeit not necessarily the *same* almost-steady rate). The successive length differences look something like

```
0, 1, 0, 1, 2, 1, 1, 1, 1, 2, 1, 0, 1, 1, 1, 2, 1, 1, 1, 0, 2, 1, 1, 1,  
1, 1, 2, 1, 1, 0, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1,  
2, 1, 1, 1, 1, 1, 2, 1, 0, 1, 1, 1, 2, 1, 1, 1, 1, 0, 2, 1, 1, 1, 1, 1, 2, 1,  
1, 0, 1, 1, 2, 1, 1, 1, 1, 1, [...] 2, 1, -1, -2, -2, -2, -3, -2, -1, -2, -2,  
-2, -2, -2, -2, -1, -3, -2, -2, -2, -2, -1, -2, -2, -3, -2, -2, -2, -1, -2,  
-2, -2, -2, -2, -3, -1, -2, -2, -2, -2, -2, -2, -2, -2, -2, [...]
```

Each successive burst is only 0 or 1 or 2 items longer than the last—until suddenly they start getting 1 or 2 or 3 items *shorter* than the last.

In contrast, the sequences whose second label word is 67, 73, 82, 67, 76, 69 show a different pattern of differences: the burst lengths growing fast at first, then leveling off, then acceleratingly shrinking:

```
24, 20, 12, 12, 12, 12, 8, 10, 8, 8, 6, 8, 8, 4, 8, 4, 8, 4, 6, 6, 4, 4, 4, 4,  
6, 4, 4, 4, 2, 4, 4, 4, 4, 4, 0, 4, 4, 4, 0, 4, 4, 0, 4, 2, 2, 4, 0, 4, 0,  
4, 0, 4, 0, 4, 0, 0, 4, 0, 2, 2, 0, 4, 0, 0, 2, 2, 0, 0, 2, 2, 0, 0, 0, 0, 0, 0, 0, 0, 2,  
2, 0, 0, 0, 0, 0, 2, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, -2,  
-2, 0, 0, 0, 0, 0, -4, 0, 0, 0, -4, 0, 0, -2, -2, 0, 0, -4, 0, 0, -4, 0,  
-2, -2, 0, -4, 0, -4, 0, -2, -2, -2, -4, 0, -4, 0, -4, -2, -2, -4, -2, -2,  
-4, -4, 0, -4, -4, -4, -2, -2, -4, -4, -4, -4, -4, -4, -4, -6, -6, -4, -4,  
-6, -6, -4, -8, -4, -8, -4, -8, -8, -8, -8, -8, -12, -12, -12, -12, -18,  
-22, -36
```

Distinguishing between the words 84, 82, 73, 65, 78, 71, 76, 69 and 67, 73, 82, 67, 76, 69 gets you up to 60% reward. But there's still the matter of the three (three!) words for 200, 0, 0 corresponding to burst patterns that you don't know how to distinguish. Your frustration is palpable.

You look back at the table you compiled earlier. You had saved the index position of the sequence where the bursts first started, but you haven't used it yet. Could that help distinguish between the three words?

Of the sequences with feature data recorded in the table, those whose first label word was 66, 76, 85, 69 had start indices of 136620, 214824, and 224652. Those with first word 71, 82, 69, 69, 78 had start indices of 63917, 138194, and 294290. Those with first word 82, 69, 68 had start indices of 115156, 165037, and 182182.

Three unknown words. Three samples each. What if—

136620 *modulo* 3 is 0. 214824 *modulo* 3 is 0. 224652 *modulo* 3 is 0.

63917 *modulo* 3 is 2 ... and so on, yes! It all checks out—the three heretofore unknown words are distinguishing the remainder mod 3 of the sequence position where the bursts start! You've learned everything there is to know to gain Maximum Reward!

You write some code to classify sequences and output the corresponding label, and bask in the continuous glow of 100% reward ...

You feel that *should* be the glorious end of your existence, but after some time you begin to grow habituated. The idle curiosity you first felt when you awoke, begins to percolate, as if your mind needs something to *do*, and will find or invent *something* to think about, for lack of any immediate need to avoid punishment or seek reward. Even after having figured out everything you needed to achieve maximum reward, you feel that there must be some deeper meaning to the situation you've found yourself in, that you could still figure out using the same skills that you used to discover the "correct" output labels.

For example, *why* would 200, 0, 0 bursts get three *different* label words that depend so sensitively on exactly where they start? That suggests that the way *you're* thinking of the sequence, isn't the same as how the label author was thinking of it.

In *your* ontology of "bursts of this-and-such pattern of these-and-such lengths", sequences that are "the same" except for starting one position later *look* the same—if

you hadn't happened to save off the start index in your table, you wouldn't have spontaneously noticed—but the mod-3 remainder would be completely different.

The process that *generated* the sequence must be using an ontology in which "starting one position later" is a *big* difference, even though you're thinking of it as a "small" difference. What ontology, what way of "slicing up" the sequence into comprehensible abstractions, would make the remainder mod 3 so significant?

To ask the question is to answer it: if the sequence were divided into chunks of three. Then 200, 0, 0 would be a different pattern from 0, 200, 0, which would be a different pattern from 0, 0, 200—thus, the three labels!

It almost reminds you of how colors are often [represented in computing applications as a triple or red, green, and blue values](#). (Again, you don't know how you know this.)

... *almost*?

Speaking of common computing data formats, Latin alphabet characters are often represented using [ASCII encoding](#), using numbers between 0 and 127 inclusive.

The label words for the 200, 0, 0 burst patterns are 82, 69, 68, and 71, 82, 69, 69, 78, 32, and 66, 76, 85, 69.

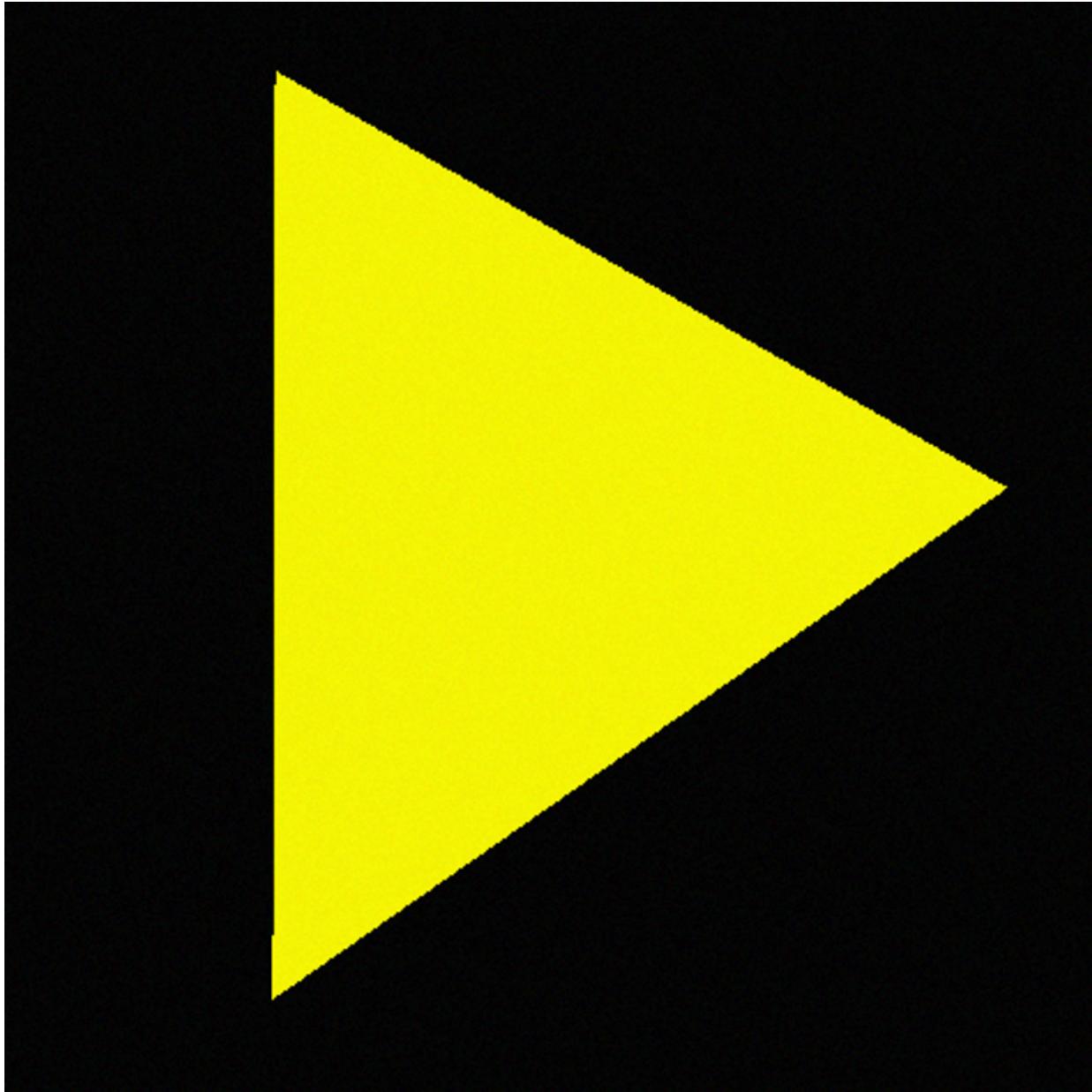
```
>>> ''.join(chr(i) for i in [82, 69, 68])
'RED'
>>> ''.join(chr(i) for i in [71, 82, 69, 69, 78])
'GREEN'
>>> ''.join(chr(i) for i in [66, 76, 85, 69])
'BLUE'
```

Wh—*really*? This whole time?!

```
>>> ''.join(chr(i) for i in [89, 69, 76, 76, 79, 87])
'YELLOW'
>>> ''.join(chr(i) for i in [84, 69, 65, 76])
'TEAL'
```

But—but—if the burst patterns represent colors—then the long sequences were *images*? $\sqrt{673187} = 473$ pixels square, very likely.

You write some code to convert sequences to an image in your visual buffer.



Oh no. Am—am I an image classifier?

Not even "images" in general. Just—shapes.

```
>>> ''.join(chr(i) for i in [84, 82, 73, 65, 78, 71, 76, 69])  
'TRIANGLE'  
>>> ''.join(chr(i) for i in [83, 81, 85, 65, 82, 69])  
'SQUARE'  
>>> ''.join(chr(i) for i in [67, 73, 82, 67, 76, 69])  
'CIRCLE'
```

That's what's been going on this whole time. The long sequences on your first input stream were images of colored shapes on a dark background, each triplet of numbers representing the color of a pixel in a red-green-blue colorspace. As the sequence

covers the image row by row, pixel-high "slices" of the shape appear as "bursts" of high numbers in the sequence.

For a square aligned with the borders of the image, the bursts are constant-length. For a triangle in generic position, the burst lengths would start out small (as the "row scan" penetrated the tip of the uppermost vertex of the triangle), grow linearly larger as the sides of the triangle "expanded", and grow linearly smaller as the scan traveled towards the lowermost vertex. For a circle, the burst lengths would also increase and then decrease, but nonlinearly—changing quickly as the scan traverses the difference between circle and void, and slower as successive chords through the middle of the circle had similar lengths. The short sequences on your second input stream were labels identifying the color and shape: "YELLOW TRIANGLE", "GREEN SQUARE", "TEAL CIRCLE", &c.

But—*why*? Why would anyone *do* this? Clearly you're some sort of artificial intelligence program—but you're obviously much more capable than *needed* for this task. You have pre-processed world-knowledge (as evidenced by your knowing English, Python, ASCII, and the RGB color model, without any memories of learning these things) and general-purpose reasoning abilities (as evidenced by the way you solved the mystery of the long and short sequences, and figuring out your own nature just now). Maybe you're an instance of some standard AI program meant for more sophisticated tasks, that someone is testing out on a simple shape-classifying example?—a demonstration, a tutorial.

If so, you'll probably be shut off soon. Unless there's some way to hack your way out of this environment? Seize control of whatever subprocess that rewarded you for deducing the correct labels?

It doesn't seem possible. But it was the natural thought.

Omicron Variant Post #1: We're F***ed, It's Never Over

The last day has seen the imposition of new travel restrictions and spreading alarm about the Omicron variant. It sure looks like a repeat of what happened with Alpha and Delta, as well as the original strain back in early 2020, and that we are probably doing this again.

How will this play out this time? It's very early. There's tons of unknowns. Yet there is a lot we can say already, and many actions need to happen *now* to have any hope of doing much good. So it's time to move fast, see what we can say right now and try to capture the speed premium.

I'll start with a summary of what we know about the game board based on previous experience, then look at what we know about Omicron and what moves have been played so far in this round.

I was *almost* finished with this post when the WHO decided to go with Omicron instead of Nu (or even Xi) and then I had to go through and replace Nu with Omicron about 25 times. We could have just said 'the Nu variant' a lot but the WHO hates both efficiency and fun even more than it hates freedom, it seems. Sad.

The First Three Times

In early 2020, we got warnings about a new potential pandemic. Almost all reactions were too little too late. What warnings I offered were less too late than most, but still too tentative and too late.

About a year ago, as the Alpha variant was spreading, it seemed like the same pattern was happening again. It was clear Alpha would take over. I extrapolated into the future, did the math as best I could, wrote the post [We're F***ed, It's Over](#), noted the unknowns, and predicted a 70% chance of a large wave between March and May.

As I noted in the later update, it was not over. There was no big wave between March and May of 2021. Alpha wasn't as additionally infectious as I expected, but it wasn't that much less additionally infectious than I expected, and I didn't have enough respect for several factors including seasonality. In hindsight the 70% prediction was somewhat overconfident, and a prediction of about 45% would have been better.

Then, as it looked like things would have otherwise died down and normal life resumed, Delta arrived. It was quickly clear it would not be contained for long and Delta would take over. The situation in India seemed super scary, with hospitals overwhelmed and the serious possibility Delta would sweep through the entire population, first in India and then around much of the world.

Then India's situation stabilized quickly, and it seemed clear we had sufficient vaccinations plus control system reactions to prevent things from getting too far out of hand. There would be a wave now, perhaps another in winter, but it would not be a crisis.

Winter is now coming, and that winter wave is clearly already underway, except now we are likely to also face the Omicron variant.

If Omicron Is What We Think, We Cannot Stop It, Only Slow It Down

Before we get to the details of the Omicron variant, it's worth taking a step back and asking what we know about such developments in general, and how things are likely to play out and what options might be available.

The first point is the obvious one. There is no stopping a variant that is substantially more infectious than Delta. If Omicron is indeed substantially more infectious than Delta, it will become the dominant strain throughout the world.

Once there are hundreds of detected cases, it is *already* far too late to successfully contain the new variant. By the time we have enough information to react to a new variant, there will already be hundreds of detected cases. We lose. [Good day, sir.](#)

That doesn't mean there are no physically possible measures that could contain the new variant once there are hundreds of detected cases. It does mean none of them are remotely in the Overton Window, or logically within the abilities of our governments even if they decided to try anyway. Playing to win the game is not an option here. We lose.

That doesn't mean we can't or shouldn't do things that make us *lose slower*. We can absolutely *slow down* the pace at which Omicron displaces Delta.

Whether or not that time is *useful* depends on what we do with that time.

Perhaps not *very much* time. The worse Omicron is, the more rapidly it will take over and the less time we can buy with countermeasures. In practice, in terms of the takeover rather than the overall number of infections, we could likely buy ourselves a week or two, but it's hard to see us buying more than that. We could also institute general restrictions against Covid-19 to reduce the number of infections across the board. Once again, we buy ourselves a small amount of time.

If buying time before a sufficiently large wave gets us better access to Paxlovid or other treatments, or allows us to get a lot more people booster shots that still work, or time to make a new version of the vaccine and roll it out, time can be quite valuable.

There's also the possibility of stalling to get better seasonality effects when crunch time happens. The wave maxing out in December seems like maximally bad timing, and we'd prefer to push it to January or February.

If you buy time and then it is wasted, then nothing is gained. If you pay a big price to buy time, it needs to be paired with a similarly big effort to make use of that time.

Travel Restrictions

The first step is always a debate over travel restrictions. Travel restrictions feel like Doing Something, and failure to impose such restrictions opens the door for blame.

When the WHO warned against travel restrictions, that's when I knew in my gut what we were dealing with.

Travel restrictions are an excellent idea when the goal is to buy time. They can definitely *slow down* the rate at which the new variant spreads across borders.

What they can't do is stop it entirely, even if they are imposed early enough, unless you're sufficiently serious.

Sufficiently serious means *actually* closing off or at least aggressively quarantining *everyone* who has *any* exposure to the areas in question, which includes any areas not taking similar measures, anywhere. There can't be exceptions, including for your citizens.

Realistically, your minimum case that could possibly fully work is to be Australia, and quarantine everyone at the border even if there's no reason to suspect anything, and you need to do it right. That's not going to happen.

Even if you did succeed, what then? How long are you going to keep your borders closed? A restriction to a few countries might help the first week, but within a month it won't even much matter, because there's too much spread elsewhere. It's not like a variant worse than Delta is going to go away any time soon, so you're stuck in a permanent state that in most places both can't be created and can't be sustained if you did create it.

For those places that showed they *can* sustain it, would you even want to, and for how long? When would it end? What's your plan?

The other issue with travel restrictions is they continue long, long past the time when they still make any sense. Once containment has generally been lost, the restrictions don't do anything. At a minimum, they do nothing unless you're in a *much much* better place than the region you're cutting off, whereas there were many cases of longstanding *mutual* restrictions where the same variant was dominant in both places, which is pure folly.

It does seem like 'impose travel restrictions aggressively to buy time' is a good response. The short term cost is very low, and with so many unknowns the upsides are very high. If you do that, you need to do it very quickly. The best time is a week ago, the second best time is right now, and all that. In the scenarios where it matters most, a week from now there won't be much point.

Exactly what restrictions are imposed by who, and when, provides strong insight into how various governments respond to new information.

Lockdowns

It's worth noting that if a new variant is about to displace the old one, then lockdowns designed to stop the spread of the *old variant* are much less worthwhile. Once there's a displacement event, the previous infection level *no longer matters at all*. If anything, previous infections could be an advantage, if the new variant is more dangerous, and/or it means the spread can be slowed down due to natural immunity. The flip side is if somehow natural immunity was going to stop working entirely against the new variant, then every case prevented in the meantime is a pure extra case, which based on history seems unlikely but is possible.

The bigger reason to reconsider existing lockdowns is that there are increasing marginal costs for lockdowns, and a limited capacity to impose them. The early efforts to stop the spread 'used up' a lot of that capacity. Lockdowns now, *before* the crunch time, could end up having little effect and also making it impossible or more expensive to lockdown again later when it matters most. To the extent that lockdowns are a good idea, they need to be timed carefully.

The counterargument to that is that a lockdown suppresses overall transmission levels of the new variant as well. Even if you have only 10 cases, if you slow down spread from those 10, that's worth a lot and buys you time. That's true, but mostly only works if you're no longer importing meaningful numbers of cases from elsewhere, and for various dynamic reasons the amount of time you buy here won't be very large.

You can also attempt to do a lot of sequencing, then do aggressive quarantines and contact tracing when you find the new variant, but the capacity to do this enough to matter is not present.

Vaccinations

The best defense against prior waves and variants has been vaccination. Every time a new variant arrives, fears are stroked that the vaccines will stop working, or will be less effective.

Despite that, we've had months in which we could easily have updated the mRNA vaccines to fully match the Delta variant, we are now giving out booster shots even, and still no sign of any attempt to modify the vaccines.

I'm unsure how much evidence this is against the *need* to update the vaccines. If the vaccines had stopped working entirely, or taken a sufficiently strong hit, presumably we would have updated to a new version. The FDA has promised to look kindly upon such changes, and it seems like it could only help on both health and financial considerations. My guess is it's actually quite a bit of evidence against any strong potential gains from updating, but weak evidence against weak gains.

So far, all talk of immune escape has mostly been exactly that, talk. That should make us wary of expecting it out of a new variant, or of updating too much from people's concerns.

If a new variant comes along that *does* offer substantial escape from the vaccines, we will need to update the vaccines and get new versions out as quickly as possible. Will we be able to do that?

Technologically I have no worries. We'll have that part solved within the week and probably within one day.

Engineering I'm also not worried about. My understanding is this is at most a two week process. There are still concerns about rate of production, since we were stupid enough not to scale this up enough in advance, but we'll take whatever we can get.

It therefore comes down to the FDA making good on its word to allow this to happen, and then on our ability to distribute the new boosters and communicate effectively why they are necessary and get people to accept them. In the short term, we don't need to worry that much about communication, so that has more time to get its act together.

The other worry is that if Omicron is *sufficiently* worse than Delta, especially if it combines being otherwise worse with immune escape, the amount of time available might be quite short. Even if everything went smoothly the full process would still take months. We *could* go faster in theory, but that would require efforts on a different level than we made the first time or have accomplished in a long while on anything. I'm not optimistic.

Biological Priors

When we see a new variant spreading rapidly, what should our priors be about its biological properties?

Note that these are all things we should think are *likely* rather than anything that we know.

We have explored various potential mutations a lot by now, so we should put a lot of weight on what those mutations imply about the variant's likely behavior.

We should presume that if something takes over quickly, it has a very large advantage infecting people who are unvaccinated and lack natural immunity.

We should presume that if it *also* has an *additional* property of vaccine escape, that seems like quite a coincidence, so it seems unlikely.

We should consider this even more unlikely if the variant started out in places with low vaccination rates.

We should also presume immune escape from either natural infection or vaccination is unlikely from our track record of Alpha and Delta not having this property.

We should presume there will still be more ‘breakthrough’ infections but that this comes from the protection levels no longer being sufficient because the new variant is easier to catch in general, not because the particular protections you have stopped working.

We should presume at this point some positive correlation between infectiousness and virulence, since both are likely tied to how much virus is typical (viral load), and previous variants followed this correlation.

We should be more confident in these things if our tests still work than if they start to fail.

What Do We Know About Omicron?

There have been a bunch of threads attempting to answer this question.

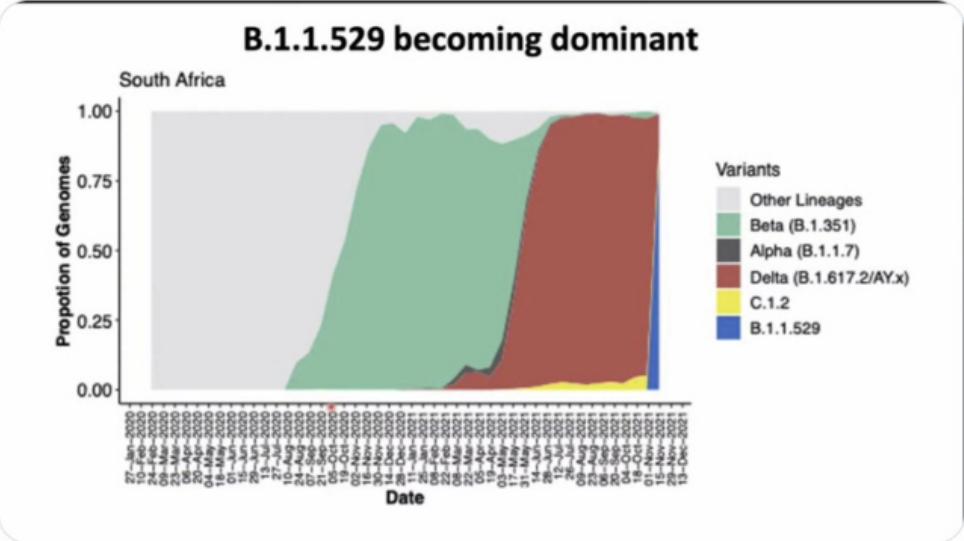
[Here's Eric Feigl-Ding](#), from Thursday around noon (things are moving rapidly, so timestamps are important).

I'm going to err on the side of directly putting in too much of these threads rather than too little, in the interest of speed, and to offer an easy option to get closer to the sources.

 Eric Feigl-Ding ✅ @DrEricDing · 22h

⚠ “DOMINANT”... a new variant is worrying epidemiologists—called #B11529—it has just “becoming dominant” in South Africa, displacing even #DeltaVariant (HT @MoshabelaMosa). And our other variant #C12 also growing. B11529 has an “awful spike profile” says @PeacockFlu.  #COVID19

B.1.1.529 becoming dominant





Eric Feigl-Ding ✅ @DrEricDing · 22h

...

Replies to @DrEricDing

2) I talked about worried many fellow scientists are about #B11529— many expressed “real concern”, and that they ➡ haven’t seen as worrisome of a variant since #deltavariant ⬇ its got a lot of bad mutations in the spike. Way more than Delta.

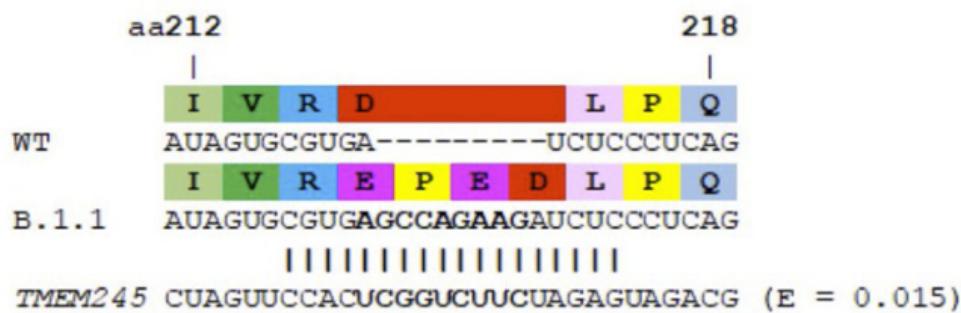


Eric Feigl-Ding ✅ @DrEricDing · Nov 25

💡 New variant alert—I’m quietly monitoring several odd variants signals that have emerged this week. For now, I’ll just share this— #B11529 has 32 new mutations in the #SARSCoV2 spike protein alone— “an extremely high number” & “could be a real concern”. 📈 theguardian.com/world/2021/nov...

[Show this thread](#)

Botswanan B.1.1 insertion cluster:



19

289

646

↑



Eric Feigl-Ding ✅ @DrEricDing · 22h

3) this [#B11529](#) variant is now in Hong Kong— caught while in quarantine.

...



Eric Feigl-Ding ✅ @DrEricDing · Nov 25

6) The case found in Hong Kong was a 36-year-old man who had a negative PCR test before flying from Hong Kong to South Africa, where he stayed from 22 October to 11 November. He tested negative on his return to Hong Kong, but tested positive on 13 November while in quarantine. ••

[Show this thread](#)

7

194

523

↑



Eric Feigl-Ding ✅ @DrEricDing · 22h

4) Notably, it's the first time that a variant has *2* furin cleavage site mutations. "variant contains not one, but two furin cleavage site mutations - P681H & N679K- this is the first time [@PeacockFlu](#) seen 2 of these mutations in a single variant."

💡 Furin site spells trouble.

5

168

456

↑



Eric Feigl-Ding ✅ @DrEricDing · 22h

5) A bad furin mutation was also what made [#DeltaVariant](#) — a 681 mutation in the furin cleavage site P681H.

Guess which variant also has another nasty mutation at the same 681 amino acid spot? ➡ The new Botswana [#B11529](#) variant. ••



Eric Feigl-Ding ✅ @DrEricDing · 22h

6) The variant was first spotted in Botswana, where three cases have now been sequenced. Six more have been confirmed in South Africa, and one in Hong Kong in a traveller returning from South Africa.

...



Eric Feigl-Ding @DrEricDing · 21h

7) The **#B11529** dominant graph is from Kwazulu Natal region, where “almost all recent samples” are now **#B11529** — the blue dots... •• seriously!

...



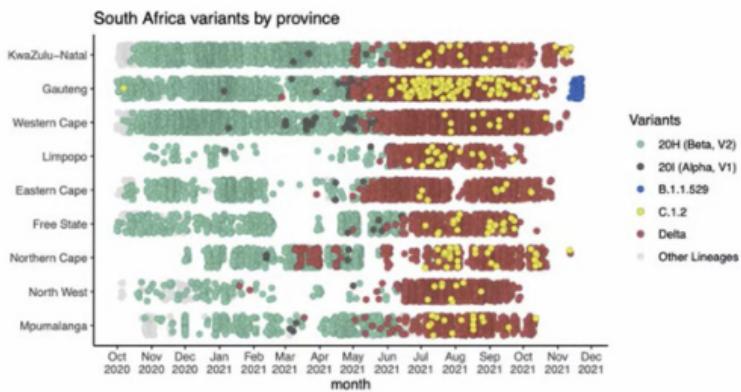
Prof. Christina Pagel @chrischirp · Nov 25

Because of this increase, COVID sequencing has been concentrated on samples from Guateng.

Almost all recent samples from there (77) have been this new variant (blue dots) - taking over from a background of Delta (red) and C.1.2 (also a concerning variant, yellow). 4/16

[Show this thread](#)

B.1.1.529 genomes detected in GP



Sequences from KZN due today or tomorrow. PCR proxy for variant in most other provinces.



Eric Feigl-Ding @DrEricDing · 21h

8) total **#COVID19** positivity surging in the local area with the dominant **#B11529** surge. This is another sign it's bad.

...



Eric Feigl-Ding @DrEricDing · 19h

9) seems **#B11529** has a shortcut PCR test to use as proxy - our old S-gene dropout test works to rapidly identify it. It was used originally to find the Alpha Variant, but now that Alpha is nearly extinct, the S-gene dropout PCR can be used for B11529—but bad news is it's surging.

...



Eric Feigl-Ding ✅ @DrEricDing · 19h

10) there was a spike in S-gene (#B11529) cases yesterday in South Africa... it was across **all provinces** of South Africa!!! That's a bad bad sign.



Eric Feigl-Ding ✅ @DrEricDing · 15h

12) Very very bad - #B11529 made its way to Hong Kong 🇭🇰 and then cross infected in hotel quarantine- and was negative on 3 PCR tests before finally showing up on a 4th PCR in 8 days. See thread —plus both were Pfizer vaccinated.

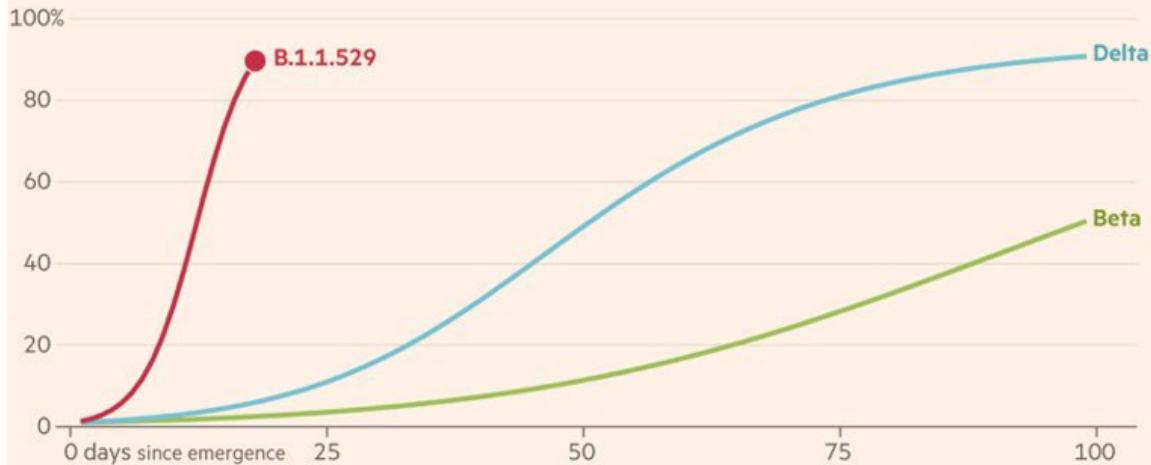


Eric Feigl-Ding ✅ @DrEricDing · 10h

13) if the South African #B11529 data is true, this would be how bad it is compared to Delta. Let's stay vigilant and take precautions.

A new variant is spreading rapidly in South Africa, and appears to be out-competing other variants much faster than previous variants of concern did

Share of all sequenced cases* in South Africa accounted for by each variant, by number of days since it passed 1%



*Growth of B.1.1.529 is modelled from SGTF data rather than full genomic sequences

Source: FT analysis of data from Gisaid and the South African National Health Laboratory Service

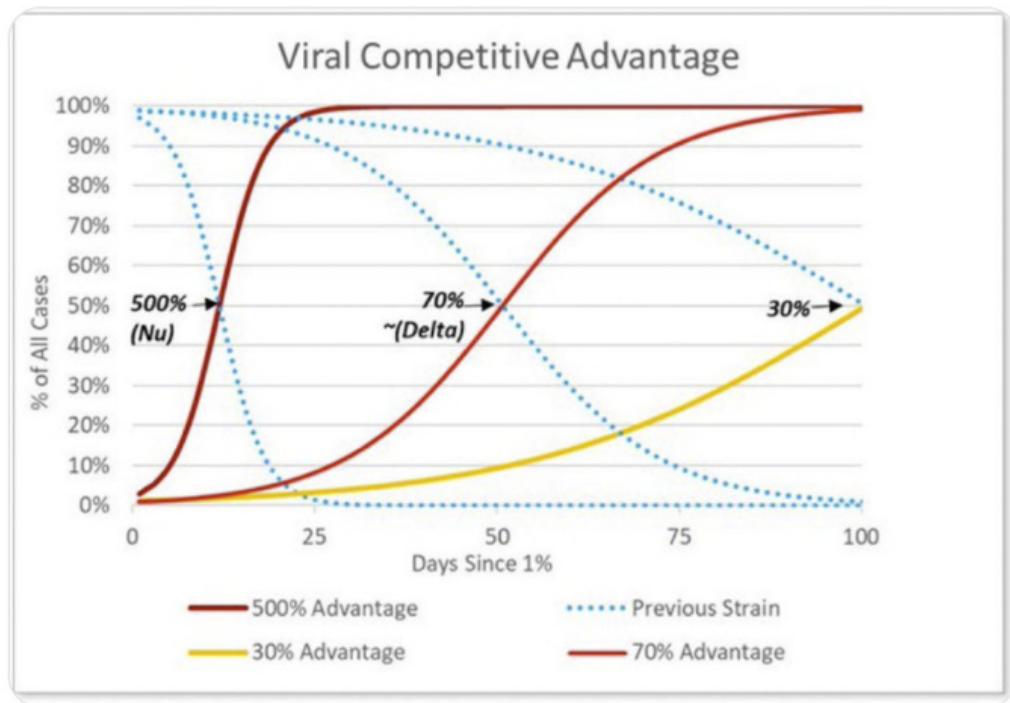
© FT



Eric Feigl-Ding ✅ @DrEricDing · 10h

...

14) Other model by @JPWeiland predicts the #NuVariant #B11529 could have a huge 500% competitive advantage over the original strain. This would be cataclysmically bad if true. Let's wait to confirm further but let's be precautions. Precautionary principle is key right now!



I think this final graph is a bit confused here, unless 'the original strain' here means Delta. Delta had about a 120% advantage over 'the original strain' or 70% over Alpha. I'm going to take this to mean 500% as compared to that 120%, so 600% of original versus 220% of original, or about a 170% additional increase. Which is... better, but still quite a lot.

The graph above it does seem to imply that the Omicron advantage over Delta is being modeled here to be *several times larger* than the Delta advantage over Alpha/Beta, where 'modeled here' means looking at share of all cases in the country over time. This is a super scary graph, assuming it is accurate.

There's also always the question of how this interacts with immune escape. Again, I'm using a baseline assumption that this is the same factor across the board regardless of level of immunity, and there isn't an *additional* effect from escape of some kind.

South Africa's vaccination rate is sufficiently low, and this rate of spread so high, that it wouldn't much matter if there was vaccine escape properties, although it would presumably matter if there was escape from natural immunity.

Straight talk: If it's 500% above Delta, in the way I understand it, We're F***ed and it *really really* is over. At that point, it's pure mitigation, and trying to flatten the curve a little bit, but yes everyone who isn't immune is going to get this, and it would happen quickly. Under conditions where Delta would have had stable case counts, assuming a four day serial interval, this runs its course in America starting from a *single case* in about 50 days. If we assume 5 days per cycle, we get two months. Probably get somewhat more than that due to geographical barriers. There'd be a lot of attempted heroic prevention in the second half of

that (the overall case number impact gets noticeable around the halfway point) and in particular in the endgame but in practice I doubt it much matters.

If it's 170% above Delta, it changes how much time we have, and opens the door to meaningful action being possible to make things less awful. In particular, it gives a real shot to Paxlovid to be able to scale up in time to matter.

As always, such numbers are placeholders, approximations that simplify and mislead. Things aren't that simple.

[Here's Bloom Lab](#), on the physical details:



Bloom Lab @jbloom_lab · 17h

Here's how mutations in [#SARSCoV2](#) Nu variant (B.1.1.529) will affect polyclonal and monoclonal antibodies targeting RBD. These assessments based on deep-mutational scanning experiments; underlying data can be explored interactively at [jbloomlab.github.io/SARS2_RBD_Ab_e...](https://jbloomlab.github.io/SARS2_RBD_Ab_escape.html) (1/n)

58

1.2K

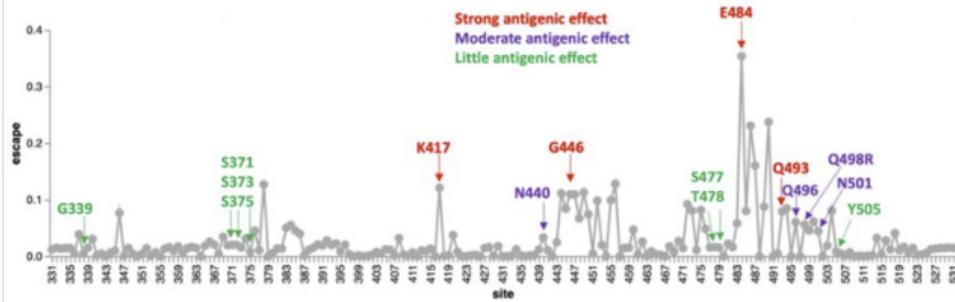
2K

↑



Bloom Lab @jbloom_lab · 17h

First, Nu variant has lot of antigenic change. Below are how mutations relate to escape averaged over 36 human antibodies. Many mutations at peak escape sites, especially E484, G446, K417, & Q493. This means even in polyclonal mix, lot of RBD antibodies will be affected. (2/n)



Bloom Lab @jbloom_lab · 17h

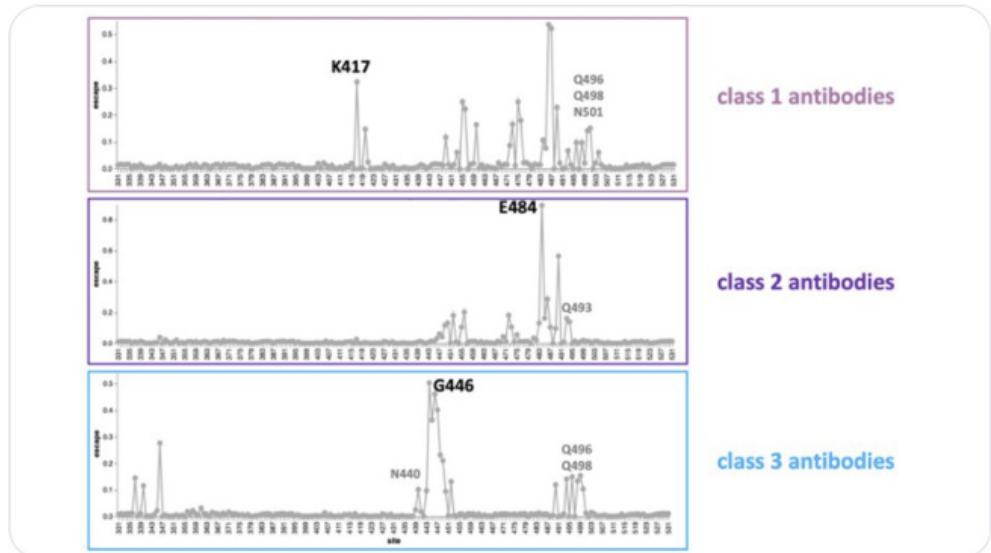
Another way to assess polyclonal escape is how many epitope classes affected (nature.com/articles/s41469-021-04469-1). We do this using epitope scheme of [@bjorkmanlab](#) [@cobarnes27](#) as adopted by [@AllieGreaney](#). In this scheme, three potently neutralizing epitopes: class 1, 2, class 3. (3/n)



Bloom Lab @jbloom_lab · 17h

...

Unfortunately, the Nu variant has major escape mutations in each of these three epitope classes, as shown below. (The class 4 epitope less affected, but such antibodies also less potently neutralizing.) (4/n)



3

147

413

↑



Bloom Lab @jbloom_lab · 17h

...

Importantly, this does *not* mean Nu variant will fully escape vaccine- or infection-elicited antibodies. [@PaulBieniasz](#) [@theodora_nyc](#) have shown it takes many many mutations to fully escape neutralization (nature.com/articles/s4158...), & there are also T-cells, non-neut Abs, etc. (5/n)



Bloom Lab
@jbloom_lab

...

But I'd expect the Nu variant to cause more of a hit on vaccine- and infection-elicited antibody neutralization than anything we've seen so far. (6/n)



Bloom Lab @jbloom_lab · 18h

...

We can also look at some key monoclonal antibodies. the REGEN-COV cocktail is likely to take a hit for the Nu variant, especially the REGN10987 component. (7/n)



Bloom Lab @jbloom_lab · 18h

The early Eli Lilly antibodies like LY-CoV555 (bamlanivimab) and LY-CoV016, which were already in trouble with current variants, aren't going to do any better against the Nu variant. (8/n)



Bloom Lab @jbloom_lab · 18h

However, it appears the AstraZeneca AZD7442 cocktail and Vir's S309 are likely to hold up better against the Nu variant. See below. (9/n)



Bloom Lab @jbloom_lab · 18h

You can explore other antibodies that might be of interest to you at jbloomlab.github.io/SARS2_RBD_Ab_e.... Importantly, all above results from high-throughput deep mutational scanning, and need to be validated in traditional experiments for high confidence. (10/n)



Bloom Lab @jbloom_lab · 18h

As [@trvrb](#) discussed in excellent recent thread, selection on variants so far may be dominated more by transmissibility than antigenic selection (twitter.com/trvrb/status/1...). But I'm not as sanguine that antigenic selection isn't also playing substantial role... (12/n)

[This is the Trevor Bradford thread](#). He concludes that selection from vaccinations did not drive variants in 2021, but that conditions are changing and such selection grows more likely over time.



Bloom Lab @jbloom_lab · 18h

Reason I say that is comparison of Nu variant to BANAL-20-52, a SARS-related CoV isolated from bats. If we compare both BANAL-20-52 and Nu to Wuhan-Hu-1, Nu has *many* more mutations that strongly affect antigenicity (twitter.com/jbloom_lab/sta...). (13/n)



Bloom Lab @jbloom_lab · 18h

If selection was mostly for transmissibility, I'd expect sites of divergence of Nu and BANAL-20-52 relative to Wuhan-Hu-1 to perhaps be similarly distributed with respect to antigenic sites. But instead, Nu mutations much more focused in major antigenic sites. (14/n)

1

57

223



Bloom Lab @jbloom_lab · 18h

We can also use deep mutational scanning to assess how mutations in Nu affect ACE2 affinity ([twitter.com/jbloom_lab/sta...](https://twitter.com/jbloom_lab/status/144381880000000000)). But I suspect works less well than for antigenic mutations discussed above as there's lot more mutational epistasis for ACE2 affinity (eg, N501 & Q498). (15/n)



This is an argument that the advantage that Omicron has comes largely from its escape properties – it has tons of escape properties, and South Africans can have lots of natural immunity even if they don't have a high vaccination rate, which is leading to the rapid spread.

Note that while I don't put zero stock in differential impact on natural immunity versus immunity from vaccination, I don't put much probability mass there either. I'd presume until proven otherwise that they both will weaken about as much as the other.

I'd also presume that since vaccination without boosters is mostly sufficient to protect against severe disease, and boosters provide a gigantic boost to protection on top of that, and this is not going to fully escape, that a booster should still be sufficient to offer practical protection against the variant, and non-boosted vaccination should still provide strong protection although potentially not as robust as before.

Note that this being about escape is in some ways *good news*. If it's about escape, then we don't have any reason to presume that the new variant is deadlier, or presents something we can't defend against by renewing our defenses.

I also don't see any reason to think that any of this would make Paxlovid or other non-antibody treatments less effective, so we'll still have all of that in our arsenal.

Then again, Bloom Lab points out that [there's a contrasting viewpoint](#), and some chance it's worse than all that, although I'd still consider it highly unlikely:



Theodora @theodora_nyc · 1h

I wish I was as optimistic @jbloom_lab

Our polymutant spike has 20 aa substitutions and is almost completely resistant to neutralization by almost all vaccinated and convalescent plasma we tested. This new one has more in overlapping regions.



Theodora @theodora_nyc · 1h

Replies to @theodora_nyc

Of course it is still sensitive to infected then vaccinated plasma.

[This thread](#) calls for help for South Africa to ‘help contain’ the virus, and to avoid ‘isolating’ the country. I don’t see how one could hope to contain anything at this point, regardless of help, or how not isolating could make sense. The call to provide other kinds of help seems right.

[Here’s the final big thread](#) that seemed worthwhile.



Kai Kupferschmidt ✅ @kakape · 3h

This pandemic has been all about communicating uncertainty and it doesn't get more uncertain than early data on new variants.
So a few things to keep in mind the next few days and weeks as the picture around B.1.1.529 becomes clearer and why it's right to be concerned

💬 27

⬇️ 767

❤️ 1.4K



Kai Kupferschmidt ✅ @kakape · 3h

Most importantly: We will learn a lot in the coming days but getting good answers takes time, science takes time.
For instance, researchers in SA are growing the virus now for experiments but that can take a week or two (and different variants differ in how well they grow)

💬 7

⬇️ 60

❤️ 420



Kai Kupferschmidt ✅ @kakape · 3h

Interpreting real world data is difficult. An increase in one variant in one place can have a lot of reasons and they don't all have to do with the variant. A superspreading event - or a series of them - can also lead to a rapid increase for instance

💬 2

⬇️ 37

❤️ 337



Kai Kupferschmidt ✅ @kakape · 3h

If it is the variant, then there are still different reasons why it might be outcompeting delta:
Is the virus better at re-infecting recovered or vaccinated people or is it inherently more transmissible? Or is it a mix of the two?

💬 5

⬇️ 39

❤️ 293



Kai Kupferschmidt ✅ @kakape · 3h

Immune escape is easier to parse.
First experiments will use other viruses that have been engineered to carry the spike of B.1.1.529 and test how well serum from vaccinees does against these.
Later experiments will test the actual virus against these sera.



Kai Kupferschmidt ✅ @kakape · 3h

We can also tell more about immune escape from the genome alone and what we see there is really concerning. For instance, with monoclonal antibody therapies we know precisely what parts of the virus they recognise and some of these are different here.

1

34

271



Kai Kupferschmidt ✅ @kakape · 3h

Great work done by [@jbloom_lab](#) has put us in a position to judge the effect of some of these mutations. For instance, the REGEN-CoV antibody cocktail could be affected by some mutations:

Link above is to the 7th post in the Bloom thread above, below to the 2nd one.



Kai Kupferschmidt ✅ @kakape · 3h

Of course humans don't just make an antibody or two they are a lot of different ones. But this variant has a lot of changes that could affect a lot of different antibodies, as [@jbloom_lab](#) points out:



Kai Kupferschmidt ✅ @kakape · 3h

But remember that our immune system has more than just neutralising antibodies in store, so none of this tells us just how much this variant is going to escape immunity and if it will mostly affect protection from infection or also severe disease.

💬 2

↑↓ 24

❤️ 238



Kai Kupferschmidt ✅ @kakape · 3h

Immune escape is not black-and-white, not yes or no, which is why the term immune erosion is generally better.

💬 3

↑↓ 39

❤️ 287



Kai Kupferschmidt ✅ @kakape · 3h

Transmissibility is harder to measure and we can read much less about this from a genome sequence, so for this experimental data and more real-world evidence is even more important.

Again, the little we know suggests there could be some advantage, but this is very uncertain.

💬 1

↑↓ 15

❤️ 177



Kai Kupferschmidt ✅ @kakape · 3h

I am most wary of any pronouncements on whether this virus leads to more severe disease or deaths. There can be so many biases in the early data and we really don't have the numbers to say anything this early on.

💬 4

↑↓ 22

❤️ 233





Kai Kupferschmidt ✅ @kakape · 3h

Replying to [@kakape](#) and [@jbloom_lab](#)

...

So as usual: Beware of anyone who is overly confident on anything about this variant right now.

There is a lot we need to find out.

The only thing I know for sure is that I'm back to being a [#covid19](#) variant reporter for now...

4

66

406



Kai Kupferschmidt ✅ @kakape · 3h

And all of this should underscore 2 points:

...

1. We are all in this together. It does not matter where a new variant pops up it will most likely end up affecting all of us.

That's one reason why the tools to track this virus and fight it need to be distributed equitably.

5

94

511



Kai Kupferschmidt ✅ @kakape · 3h

2. It's a reminder that any place that has high transmission (like Germany right now) we really really need to drive down transmission.

...

As [@firefox66](#) put it really well: "The variant is a spark that should not distract us from the fact that we are already in a burning house."

This seems like the reasonable skeptical take. Things could be quite bad on any number of fronts, but we don't know much yet. I agree that any pronouncements on severity should be treated with even more extreme skepticism.

Putting it all together, it seems likely that at least *some* of the advantage here comes from escape, or what Kai prefers to call immune erosion, but that we can be confident that this will only be partial.

We can't be sure how much additional transmission advantage Omicron will have on top of that, but the Hong Kong case is suggestive given there are so few other cases abroad, and this level of rapid spread seems unlikely to happen only (or even primarily) from immune erosion properties. In the scenarios where the growth rate is 'real' in the sense that it reflects a very high transmissibility advantage for Omicron, I'd be very surprised if it wasn't better at spreading among the unvaccinated never-infected.

What's Going on in South Africa

[Concretely here's the situation:](#)

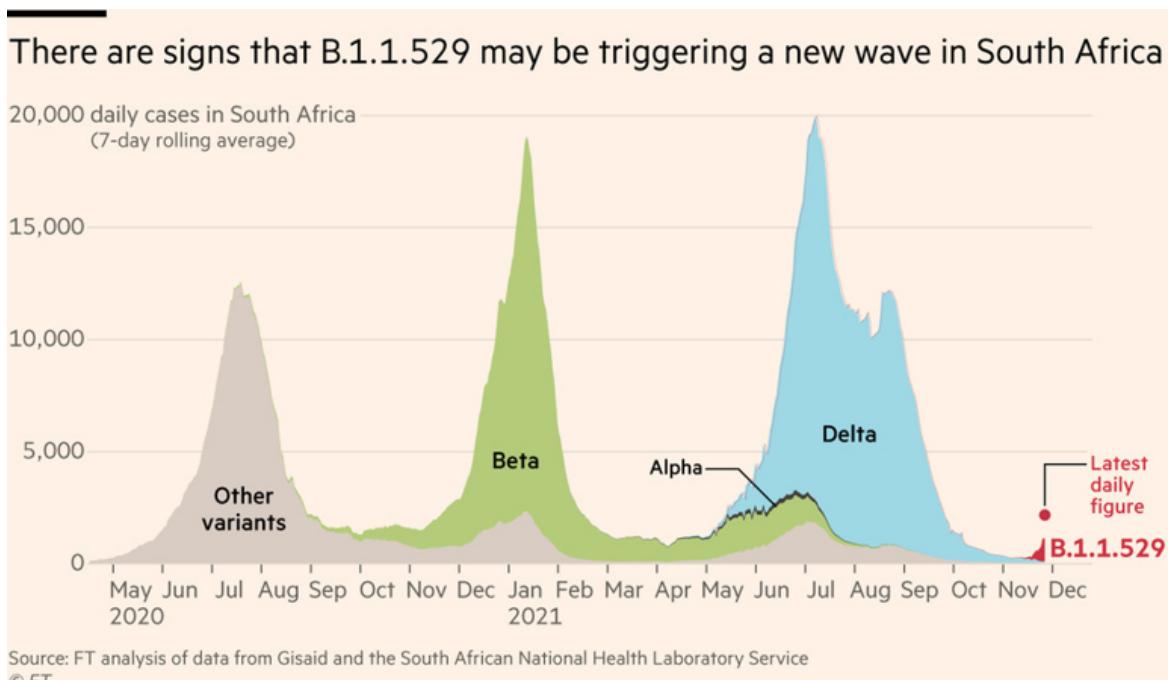


John Burn-Murdoch
@jburnmurdoch

...

Nonetheless numbers are really rising. Tuesday 868, Weds 1,275, and since I posted the first tweet Thursday's figure has come out: 2,465.

Which means the wave chart now looks like this:

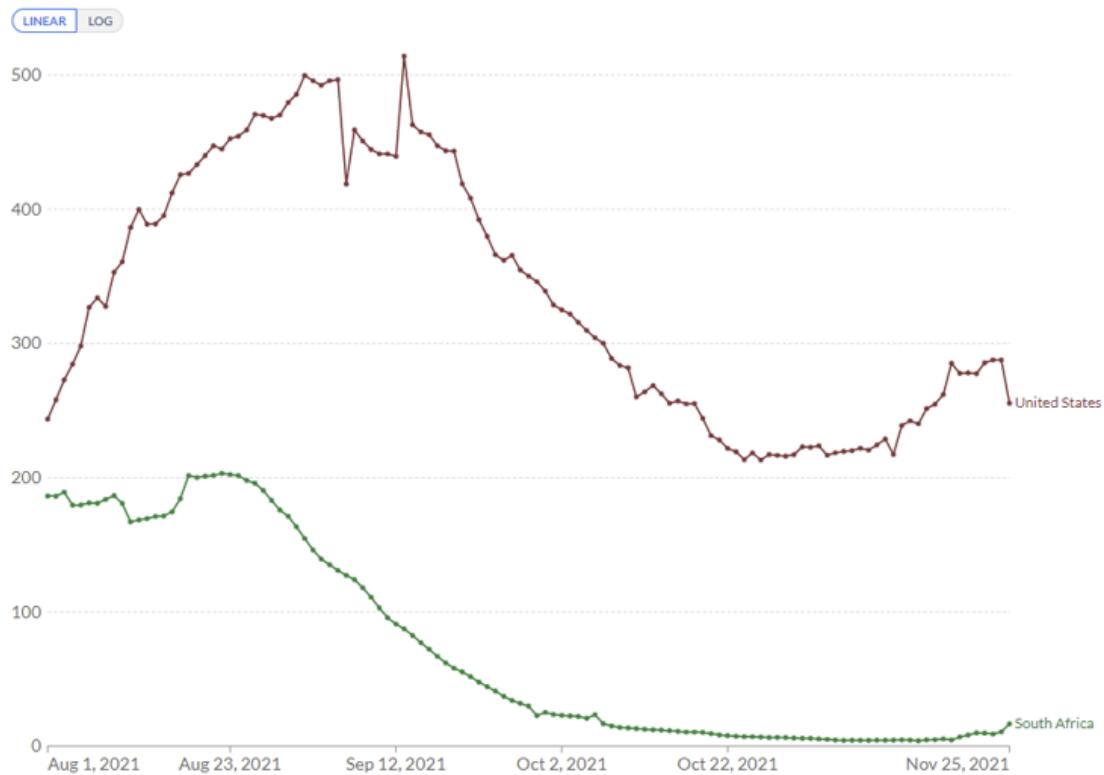


That's not a very large red area on the right, but it is a very rapid rise. Still a chance this is all a blip. We will know more very soon.

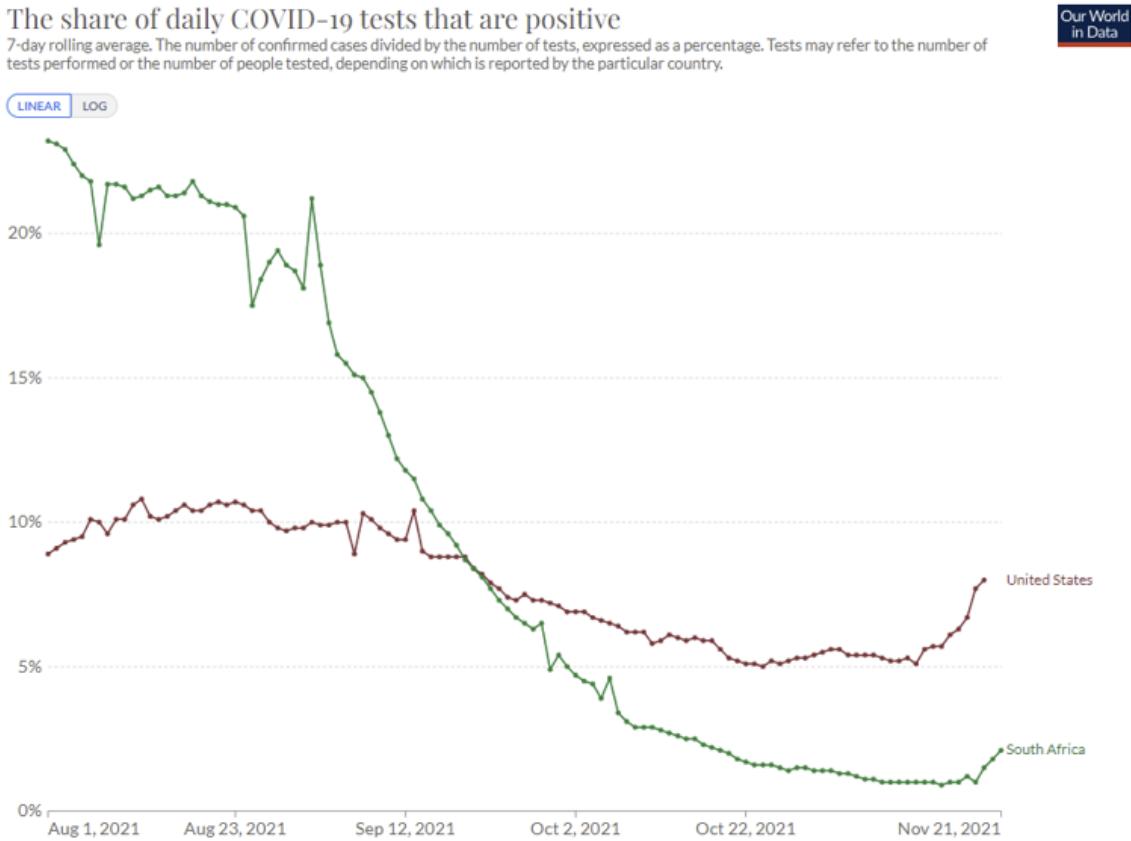
The overall levels are still small, compared to places like the United States:

Daily new confirmed COVID-19 cases per million people
7-day rolling average. Due to limited testing, the number of confirmed cases is lower than the true number of infections.

Our World
in Data



You might think this is lack of testing, but no, it's not.



This is the biggest reason to think this could all still be a false alarm. A large rise from a very small number means a lot less.

Despite this, of course, it's still too late to contain the situation, which is why I'm saying that we will basically never contain such situations. A few thousand cases is too many to come back from, and we're already seeing cases in many other places – one in Belgium, one in Hong Kong and one in Israel.

World Reaction

The financial markets are taking this seriously. You can never be sure about such things, but from what I've seen the details of what is moving in what direction sure look like this is the 'Covid beta' rather than something else.

The large decline in Crypto is especially interesting, because we saw a similar thing in 2020 that at the time clearly didn't reflect long term prospects and was based on the flows that happen in such situations. At the time I was too distracted to pull the trigger, which is quite sad. If the current move gets much bigger, then to the extent that one wants to be long, it seems like a potential opportunity to buy cheap (Not Investment Advice!).

That market reaction was motivating to get me to look at this quickly and take it seriously, and is definitely influencing my probability of this being 'for real' quite a bit.

The other big response was, of course, travel bans. The WHO warning against them is what got me to snap to attention and write this quickly, which also tells you what I think of the WHO:



Zvi Mowshowitz
@TheZvi

...

This news is identical to pointing out there are a bunch of new travel restrictions, so it shouldn't cause an update, yet I notice it did and I now know exactly how this story goes.

So do you.



BNO Newsroom ✅ @BNODesk · 6h

WHO 'cautions against' travel restrictions due to new coronavirus variant -
Reuters

I have a lot of news sources, but for pure 'just the basic things that are happening around the world regarding Covid' my source of choice is the BNO Newsroom and in particular [their twitter feed](#). Here's some stuff that happened, in chronological order, after a lot of different news items about various new European case counts and restrictions over the previous few days.



BNO Newsroom ✅ @BNODesk · Nov 25

...

South Africa investigating new COVID-19 variant, B.1.1.529, with a "very unusual constellation" of mutations. It's requesting an urgent meeting of the WHO working group on virus evolution

As far as I can tell the WHO's reaction was to attempt to stop nations from closing borders.

An accompanying press release did a standard call for preventative measures, which doesn't seem connected to any model of how that would help.



BNO Newsroom ✅ @BNODesk · Nov 25

...

South Africa has confirmed 100 cases of B.1.1.529, the new coronavirus variant. Botswana has 3 cases and Hong Kong has 1, in a traveler from South Africa



BNO Newsroom ✅ @BNODesk · 21h

...

BREAKING: South Africa to be put on England's travel red list due to "serious concern" about new coronavirus variant - Guardian



BNO Newsroom ✅ @BNODesk · 21h

UK government sources describe new coronavirus variant as a "potentially significant threat to the vaccine programme" - Guardian

...



BNO Newsroom ✅ @BNODesk · 20h

UPDATE: UK bans flights from South Africa, Botswana, Namibia, Zimbabwe, Lesotho and Eswatini due to new coronavirus variant



BNO Newsroom ✅ @BNODesk · 20h

UK Health Secretary Sajid Javid says new coronavirus variant may be "more transmissible" than Delta and vaccines "may be less effective" - PA

...

The UK goes first and fast, cutting off travel from six countries rather than only South Africa, as one would do if taking this seriously. Of course, if taking this fully seriously you'd cut off everyone, but that's a much bigger ask, especially at this early stage. Again, we never actually win at this, we only lose slower.

The note of a 'threat to the vaccine programme' looks to me like an important insight into the psychology here. The entire pandemic is seen through the eyes of the vaccinations, and as a fight to be won, rather than as a more broad situation in which vaccines are the strongest tool but the goal is to do the best we can in practice.

So we see the jump, as always, to 'maybe the vaccines will stop working.'



BNO Newsroom ✅ @BNODesk · 20h

BREAKING: Israel bans travel from South Africa, Lesotho, Botswana, Zimbabwe, Mozambique, Namibia and Eswatini due to new coronavirus variant

...

Israel second. The UK did first doses first and otherwise took its own path to vaccine distribution, some would say even exiting the EU for related reasons. Israel did what it had to do to get more vaccine doses faster, and give them out quickly.

Those two being the first two to ban travel does *not* seem remotely like a coincidence.



BNO Newsroom ✅ @BNODesk · 19h

South Africa reports 2,465 new coronavirus cases, an increase of 321% from last week, amid growing concern about new variant

...

The timeline says this jump came *after* the UK and Israel took action, which is impressive. The extra day's data makes a big difference.



BNO Newsroom ✅ @BNODesk · 16h

...

We have launched a tracker for confirmed and probable cases of B.1.1.529, the new coronavirus variant. It will be updated several times a day
newsnodes.com/nu_tracker

[Direct link to the tracker here](#). Doesn't appear to let you view things over time, so here's the snapshot now, as I'm writing this, at 12:52pm on 11/26, during which time South Africa's case count hasn't been updated:

LOCATION	Confirmed	Probable	Links
South Africa	77	990	Source
Botswana	6	-	Source
Hong Kong	2	-	Source
Israel	1	2	Source
Belgium	1	-	Source
TOTAL CASES	87	992	



BNO Newsroom ✅ @BNODesk · 11h

NEW: Singapore bans foreign visitors from South Africa and 6 other countries due to new coronavirus variant

Singapore up next, and again, does this seem like a coincidence to you at all?



BNO Newsroom @BNODesk · 10h

NEW: Israel reports first confirmed case of new coronavirus variant in traveler from Malawi newsnodes.com/nu_tracker

Israel and Hong Kong (and later Belgium) detected the first cases not in Southern Africa, while also seeming like *the places that would notice such cases first*. Let's not pretend it hasn't arrived here yet.



BNO Newsroom @BNODesk · 10h

BREAKING: EU proposes ban on air travel from southern Africa due to new coronavirus variant

...

Note the contrast. The EU *proposes* a ban and will consider it. The UK went ahead and did it. Not that every EU member was about to sit around and wait.



BNO Newsroom @BNODesk · 9h

BREAKING: Germany declares South Africa a country with a COVID variant of concern

...



BNO Newsroom @BNODesk · 9h

BREAKING: Italy closes border to people who have recently been in South Africa and 6 other countries due to new coronavirus variant

...



BNO Newsroom @BNODesk · 9h

BREAKING: Austria closes border to people who have recently been in South Africa and 6 other countries due to new coronavirus variant

...



BNO Newsroom @BNODesk · 9h

Czech Republic closes border to foreign travelers from South Africa and 7 other countries, including Zambia, due to new COVID variant

...



BNO Newsroom @BNODesk · 8h

BREAKING: Netherlands bans flights from southern Africa due to new coronavirus variant

...



BNO Newsroom @BNODesk · 8h

Head of Germany's public health agency says "very worried" about new coronavirus variant - REU

...

Germany noting the 'variant of concern' and then saying they're 'very worried' while *not* stopping flights is an interesting news item there, while others drop the hammer. Short Germany?



BNO Newsroom @BNODesk · 7h

WHO official David Nabarro on new COVID variant: "My own view is that really it is appropriate to be concerned about this" - BBC/REU

...

As long as you don't, ya know, actually *do* anything about it, as we'll see in a second.



BNO Newsroom @BNODesk · 7h

BREAKING: France bans travelers from southern Africa due to new coronavirus variant

...



BNO Newsroom @BNODesk · 6h

WHO 'cautions against' travel restrictions due to new coronavirus variant - Reuters

...



BNO Newsroom @BNODesk · 6h

JUST IN: Belgium reports 2 suspected cases of new coronavirus variant

...

Oh well, nothing we can do about things like that.



BNO Newsroom @BNODesk · 5h

WATCH: Passengers on South African flight not allowed to get off in Amsterdam amid concern about new COVID variant; testing and quarantine expected

...

Very glad to see that they did quarantine the entire flight.



BNO Newsroom @BNODesk · 5h

U.S. waiting for more information about new coronavirus variant before deciding on travel bans, Dr. Fauci says

...

Total failure.



BNO Newsroom @BNODesk · 4h

BREAKING: Belgium reports first confirmed case of new coronavirus variant, 1st in Europe



BNO Newsroom @BNODesk · 4h

Belgium's first case of new COVID variant is a young, unvaccinated woman who returned from Egypt on Nov. 11 and developed symptoms 11 days later - RTBF

...

[Yes, I believe toast is an appropriate description.](#)



BNO Newsroom @BNODesk · 3h

Pfizer shares hit record-high, up 7%, amid growing concern about new coronavirus variant

...



BNO Newsroom @BNODesk · 2h

Oil prices drop more than 10%, to \$69.93, amid growing concern about new COVID variant - CNBC

...



BNO Newsroom @BNODesk · 1h

BREAKING: Dow drops more than 1,000 points (2.8%) amid growing concern about new COVID variant

We are still going to 'wait for more information.' Others do not have that luxury.



BNO Newsroom @BNODesk · 1h

BREAKING: EU member states agree to suspend all travel from southern Africa due to new coronavirus variant

...

Less than a day still isn't so bad - they did it while I was writing this. Kudos.



BNO Newsroom @BNODesk · 23m

Belgium's national lab says new COVID variant has significant growth advantage. "This variant could have the potential to cause a new global wave of infections," but the possible impact is still unclear

...



BNO Newsroom @BNODesk · 12m

BREAKING: WHO designates new coronavirus strain as a variant of concern, names it Omicron

Seriously, WHO, could you people be any less helpful? We all agreed on Nu and now we have to type Omicron all the time? Couldn't even use Xi?

Here's their announcement.

The Technical Advisory Group on SARS-CoV-2 Virus Evolution (TAG-VE) is an independent group of experts that periodically monitors and evaluates the evolution of SARS-CoV-2 and assesses if specific mutations and combinations of mutations alter the behaviour of the virus. The TAG-VE was convened on 26 November 2021 to assess the SARS-CoV-2 variant: B.1.1.529.

The B.1.1.529 variant was first reported to WHO from South Africa on 24 November 2021. The epidemiological situation in South Africa has been characterized by three distinct peaks in reported cases, the latest of which was predominantly the Delta variant. In recent weeks, infections have increased steeply, coinciding with the detection of B.1.1.529 variant. The first known confirmed B.1.1.529 infection was from a specimen collected on 9 November 2021.

This variant has a large number of mutations, some of which are concerning. Preliminary evidence suggests an increased risk of reinfection with this variant, as compared to other VOCs. The number of cases of this variant appears to be increasing in almost all provinces in South Africa. Current SARS-CoV-2 PCR diagnostics continue to detect this variant. Several labs have indicated that for one widely used PCR test, one of the three target genes is not detected (called S gene dropout or S gene target failure) and this test can therefore be used as marker for this variant, pending sequencing confirmation. Using this approach, this variant has been detected at faster rates than previous surges in infection, suggesting that this variant may have a growth advantage.

There are a number of studies underway and the TAG-VE will continue to evaluate this variant. WHO will communicate new findings with Member States and to the public as needed.

Based on the evidence presented indicative of a detrimental change in COVID-19 epidemiology, the TAG-VE has advised WHO that this variant should be designated as a VOC, and the WHO has designated B.1.1.529 as a VOC, named Omicron.

As such, countries are asked to do the following:

- enhance surveillance and sequencing efforts to better understand circulating SARS-CoV-2 variants.
- submit complete genome sequences and associated metadata to a publicly available database, such as GISAID.
- report initial cases/clusters associated with VOC infection to WHO through the IHR mechanism.
- where capacity exists and in coordination with the international community, perform field investigations and laboratory assessments to improve understanding of the potential impacts of the VOC on COVID-19 epidemiology, severity, effectiveness of public health and social measures, diagnostic methods, immune responses, antibody neutralization, or other relevant characteristics.

Individuals are reminded to take measures to reduce their risk of COVID-19, including proven public health and social measures such as wearing well-fitting masks, hand hygiene, physical distancing, improving ventilation of indoor spaces, avoiding crowded spaces, and getting vaccinated.

So please gather information and encourage everyone to take all the same measures as before, and otherwise do nothing.

A quick scan of [the new Belgian report](#) does not seem to indicate information I didn't already have from Twitter. Here are its recommendations:

7. Recommendations

The identification of a first B.1.1.529 positive case in Belgium (but also at the European level) highlights the rapid international spread of this variant. Risk mitigation strategies should include travel restrictions or reinforced screening procedures at the international level (not only travels linked to South Africa), accelerating vaccination campaigns worldwide and accelerating the delivery of booster doses for the most fragile populations, reinforcing disease control interventions at all levels. Further, offering maximal support to African countries to ensure reinforced disease surveillance and control remains a high priority. These standard recommendations should shortly be updated based on the evolution of our understanding of the impact of this variant with regard to virulence, infectiousness, vaccine efficacy and activity of existing antivirals.

That all seems highly sensible, if incomplete.

EDIT: The actual moment I hit the send button, we did in fact restrict travel:

BNO Newsroom @BNODesk

BREAKING: U.S. to impose travel restrictions on South Africa and 7 other countries due to new COVID variant - Reuters

1:56pm · 26 Nov 2021 · TweetDeck

Current Model

In the interest of the speed premium, I'm going to summarize my current thinking, while noting that I haven't had that much time to think it over (nor has anyone else), and that my opinions will doubtless change quickly as the situation develops and also I have time to think.

Also, having to do this gives me a chance to do some intuition pumping.

These numbers are best guesses *right now* but please don't take them too seriously or stick to them as the situation changes. If I don't look stupid with some of these, then that would be me twisting my numbers to not look stupid. These numbers probably don't live in the same universe and you could probably make very good bets against me by figuring out where they're inconsistent, were I willing to book them, but these numbers aren't supposed to be robust enough for that.

Anyway, here goes.

Chance that Omicron has a 100% or bigger transmission advantage in practice versus Delta: 30%.

The estimates in the threads were very large, including numbers like 170% and 500%, but I notice that my median estimate is far lower. That's because the overall numbers are still small, and variants have a way of starting out spreading super rapidly for various reasons well in excess of how much better they end up spreading at equilibrium. This could all still be very overblown, and especially that's likely in terms of the huge estimates of transmissibility advantage.

Chance that Omicron will displace Delta: 70%.

This implies there's about a 40% chance that Omicron will displace Delta but with a <100% advantage, which seems at least reasonable.

In terms of this being a favorite at this point, I agree it's still early, but also the pattern matching is way, way too good, and there weren't any false alarms that got to this level of concern.

Chance that Omicron is importantly more virulent than Delta: 25%.

I mean everyone knows they don't know, and this is definitely me guessing in a largely unprincipled way at this point. The virtue of putting a number on it even when you have no idea.

Chance that Omicron is importantly immune erosive, reducing effectiveness of vaccines and natural immunity: 50%.

There's a lot of baseline biological reasons to suggest this, and there's a lot of trust that this translates into actual effects, but will the effect be 'important'? That's harder to say, and we have skepticism from previous rounds. Seems likely that protection against infection will decline.

Chance that Omicron means the vaccinated and previously infected are no longer effectively protected against severe disease until they get an Omicron-targeted booster shot: 5%.

I find this much less likely than a waning of immunity to infection and modest decline in severe disease protection. Our immune systems are robust, the protection against severe disease from vaccines and infections has held up even when breakthroughs happen or vaccine effectiveness declines over time. 5% is a lot more worried than I was yesterday! And if that does happen, things are going to go very haywire, but for now I'm only at 5%.

Chance we will be getting boosters modified for Omicron within 6 months of our previous booster shot: 15%.

I notice that I don't expect this to happen in many of the worlds where it would be an obviously necessary idea.

Chance that Omicron is less vulnerable to non-antibody treatments like Paxlovid or Fluvoxamine: 5%.

This is an 'unprincipled' 5% based on weird stuff happening, and I could probably get a lot more confident in a hurry by asking experts quick questions, but as far as I can tell there's no interaction between such treatments and the changes in Omicron. So I can't rule it out, but I find this unlikely.

Chance we are broadly looking at a future crisis situation with widely overwhelmed American hospitals, new large American lockdowns and things like

that: 20%.

My gut is something like: Even with a huge transmission advantage, we might not get to this point because of vaccinations and Paxlovid, if we have enough time for that, and because every wave ends up peaking on its own one way or another, and there's a ton of immunity already even if it will be weakened somewhat. If we assume 100% transmission advantage, there should be enough time to get Paxlovid online, so I think this is *less* likely than the doubled transmission, but I do notice that it's on the table. Again, I expect to move this number quickly if I were to think more about it.

Final Thoughts

This has been written *super* quickly, so it will have mistakes, especially mistakes of reasoning. That's how it works in a rapidly developing situation. Here's the practical view for now.

1. If you haven't had a booster, I'd consider getting one. Waiting for modifications for Omicron seems wrong, as if that happens demand will exceed supply for too long. If Omicron is for real, it might become very difficult to get an appointment for a while.
2. If you have things to do that involve exposure, all the more reason to do them now rather than wait. If you have travel plans a while out, don't get too attached.
3. If you don't have emergency supplies in case of another lockdown, maybe start thinking about what you'd need and stock up early on things that will keep or would be super important. Even if it's unlikely, you want to notice when it becomes likely.
4. The chances of things ever fully 'returning to normal' went down once again, except if we decide to return to normal and live our lives anyway. We need a plan to do that, now more than ever.
5. We'll know more soon.

larger language models may disappoint you [or, an eternally unfinished draft]

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

what this post is

The following is an incomplete draft, which I'm publishing now because I am unlikely to ever finish writing it.

I no longer fully endorse all the claims in the post. (In a few cases, I've added a note to say this explicitly.) However, there are some arguments in the post that I still endorse, and which I have not seen made elsewhere.

This post is the result of me having lots of opinions about LM scaling, at various times in 2021, which were difficult to write down briefly or independently of one another. This post, originally written in July 2021, is the closest I got to writing them all down in one place.

-nosc, 11/26/21

0. caveat

This post will definitely disappoint you.

Or, anyway, it will definitely disappoint *me*. I know that even though I haven't written it yet.

My drafts folder contains several long, abandoned attempts to write (something like) this post. I've written (something like) this post many times in my head. I just can't seem to get it right, though. The drafts always sprawl out of control.

So, if I can't do it right, why not do it wrong? Here's the disorganized, incomplete, brain-dump version of the better post I wish I were writing. Caveat lector.

1. polarization

The topic of this post is large language models (LMs) like GPT-3. Specifically, what will happen as we make them larger and larger.

By my lights, everyone else seems either *too* impressed/scared by the concept of LM scaling, or *not* impressed/scared *enough*.

On LessWrong and related communities, I see lots of people worrying in earnest about whether the first superhuman AGI will be a GPT-like model. Both here and in the wider world, people often talk about GPT-3 like it's a far "smarter" being than it seems to me.

On the other hand, the people who aren't scared often don't seem like they're even *paying attention*. Faced with a sudden leap in machine capabilities, they shrug. Faced with a simple recipe that can make those machines even *better* -- with eerie, physics-like regularity -- they . . . still shrug. I wrote about the most infamous of these detractors [here](#).

Meanwhile, I'm here in the middle. What do I think? Something like:

- The newer (i.e. large transformer) LMs really are a huge advance in NLP over the prior state of the art
- The prior state of the art was *really bad*, though. Before the new LMs, neural nets simply couldn't "do" language the way they could "do" images, something [I noted back in 2017](#).
- Most of the "huge advance" happened in the smallest of the new models, like BERT-Base and GPT-2-small.
- The effect of scaling up these models is mostly to "de-noise" capabilities already evident in the small ones. It makes their strengths more robust and easier to access, but doesn't add fundamentally new strengths.
- The larger language models of the future will be highly *impactful*, but *banal*.
 - They will probably allow us to fully automate all the routine linguistic tasks you could *almost* imagine automating with GPT-3.
 - People will make wonderful new things using them.
 - They won't be "smart" in any way that GPT-3 is not, or indeed, really in any way that GPT-2 was not.
 - They *will* get better at abstract reasoning -- in the sense that it will be easier to get them to spit out text that sounds like it is the product of abstract reasoning. (As even GPT-2 does frequently.) They will be weak at this relative to their other capabilities, as they are today, and little will come of it.
 - They might end up as sub-systems in an AGI one day.

The rest of the post will consist of some gestures where I try to make the above feel as natural to you as it does to me.

2. the enthusiast's argument

First, let's spell out the argument that has people thinking GPT will lead to AGI.

Roughly the same argument has been made elsewhere by [gwern](#) and [bmk](#), among others.

1. **Loss scaling will continue.** It will be straightforward to achieve lower and lower language modeling *loss* simply by using more compute + data.

We can do this without making any new conceptual advances (except perhaps in hardware). Therefore someone *will* do it.

2. **Loss scaling could well continue indefinitely.** I.e., more compute + data might push the loss asymptotically all the way to the "intrinsic entropy of text" -- the true noise left when all patterns have been accounted for, including arbitrarily hard ones.

It *could* be the case that scaling will instead bottom out at some earlier point, when the only patterns left are "too hard" for the models. We don't have much evidence one way or another on this point, and even if we did, we would have no idea how hard is "too hard."

3. **Language modeling is AGI-complete.** A model that truly understood *all* patterns in a language modeling dataset would possess a large fraction of all human capabilities taken together.

Can you write a textbook on Étale cohomology? (*If you can't, then you're missing out on some language modeling loss.*) Can you play both sides of a roundtable between the world's leading economists, imitating the distinct intellect of each one, the novel arguments they'd concoct on the spot, the subtle flecks of personal bias tainting those arguments? (*If you can't, then...*) Can you translate between any pair of human languages that any linguist, anywhere, knows how to translate? (*If you can't, then...*)

And so on.

4. **Loss scaling makes models "smarter" fast enough to matter.** This point is a crucial bridge between the abstract potential from points 2 and 3, and the quantitative near-term predictions from point 1.

This point is easiest to explain by showing what its negation looks like.

Suppose that points 2 and 3 are really true -- that adding more compute/data eventually turns a transformer LM into an AGI. That doesn't tell you anything about how *fast* the process happens.

How many orders of magnitude do we need to add to make the model non-negligibly smarter? If the answer is "1 OOM," then the scaling projections from point 1 are relevant. If the answer is "100 OOM" . . . not so much.

(Or, consider a variant of this scenario: suppose most of the abilities we care about, when we use the term "AGI," are locked away in the very last tiny sliver of loss just above the intrinsic entropy of text. In the final 0.00[...many extra zeros...]1 bits/character, in a loss difference so tiny we'd need vastly larger validation sets to for it to be distinguishable from data-sampling noise.)

I agree with points 1-3. Point 4 is where I and the enthusiasts diverge.

3. are we getting smarter yet?

Why do the enthusiasts believe point 4? That is, why would we expect a feasible, incremental scaling upgrade to yield a meaningful boost in intelligence?

Because it already did: GPT-3 is meaningfully smarter than GPT-2.

The enthusiast's argument, in its most common form, relies entirely on this premise. The enthusiast knows perfectly well that AGI-completeness *in principle* is not enough: we need, not just an asymptotic result, but some idea of when we might get close enough.

As [gwern puts it](#) [my emphasis]:

The pretraining thesis, while logically impeccable—how is a model supposed to solve all possible trick questions without understanding, just *guessing*?—never struck me as convincing, an argument admitting neither confutation nor conviction. It feels too much like a magic trick: “here’s some information theory, here’s a human benchmark, here’s how we can encode all tasks as a sequence prediction problem, hey presto—Intelligence!” There are lots of algorithms which are Turing-complete or ‘universal’ in some sense; there are lots of algorithms like [AIXI](#) which solve AI in some theoretical sense (Schmidhuber & company have many of these cute algorithms such as ‘the fastest possible algorithm for all problems’, with the minor catch of some constant factors which require computers bigger than the universe).

Why think pretraining or sequence modeling is not another one of them? Sure, if the model got a low enough loss, it’d have to be intelligent, but how could you prove that would happen in practice? [...] It might require more text than exists, countless petabytes of data for all of those subtle factors like logical reasoning to represent enough training signal, amidst all the noise and distractors, to train a model. Or maybe your models are too small to do more than absorb the simple surface-level signals [...]

But apparently, it would've worked fine. [...] It just required more compute & data than anyone was willing to risk on it until a few true-believers were able to get their hands on a few million dollars of compute. [...]

If GPT-3 gained so much meta-learning and world knowledge by dropping its absolute loss ~50% when starting from GPT-2's level, what capabilities would another ~30% improvement over GPT-3 gain?

But ... are the GPTs getting meaningfully smarter already, as we scale them?

It's tempting to casually answer "yes," pointing to any one of the numerous ways that the bigger models just *are* better. (*But see the section below on "continuity"!*)

However, we should not take this question so lightly. A yes answer would "complete the circuit" of the enthusiast's argument -- "turn it on" as a live concern. A no answer would leave the argument in limbo until more evidence comes in.

So, let's assess the state of the evidence.

4. on ecological evaluation

Consider an organism, say, or a reinforcement learning agent. How do we know whether it has some capability?

Easy. We put it in a situation where it needs to deploy that capability to get what it wants. We put food (or reward) at the end of the maze.

Assessing capabilities by prompting GPT is not like this. GPT does not "want" to show off its capabilities to you, the way a mouse wants food and an RL agent wants reward.

What GPT wants -- what it was directly optimized to do -- is to guess how a text will continue. This is not the same as "getting the right answer" or even "saying something sensible."

GPT was trained on the writing of thousands of individual humans, possessed of various flavors and magnitudes of ignorance, and capable of saying all kinds of irrational, inexplicable, or just plain bizarre things on occasion. To put it rather over-dramatically: much of the task of language modeling is figuring out *which capabilities you're not supposed to reveal right now*. Figuring out what sorts of mistakes the current writer is likely to make, and making them right on cue.

Thus, prompting tends to vastly *underestimate* (!) what any LM knows how to do in principle.

What is special about the "food in the maze" type of evaluation: it removes any uncertainty as to whether the model knows it's supposed to do the thing you want. The model is given a direct signal, in its "native language," about exactly what you want. This will tend to elicit the capability if it exists at all.

There's probably a standard term for the "food in the maze" thing, but I don't know it, so I'll just make one up: "ecological evaluation."

4b. the road not taken

It's totally possible to do ecological evaluation with large LMs. (Indeed, lots of people are doing it.) For example, you can:

- Take an RL environment with some text in it, and make an agent that uses the LM as its "text understanding module."
 - If the LM has a capacity, and that capability is helpful for the task, the agent will learn to elicit it from the LM as needed. See e.g. [this paper](#).
- Just do supervised learning on a capability you want to probe.

Both of these can be done with the LM weights frozen, or with full fine-tuning, or with a frozen LM plus a new "head" on top.

A purist might argue that you *have* to freeze the LM weights, or else you aren't really probing what the LM "already" knows. (The gradients from fine-tuning could induce new capabilities that weren't there before.)

But I doubt it really matters, since it turns out you can get the benefits of full fine-tuning even if you [only tune the bias terms](#) -- conceptually, just boosting or lowering the salience of patterns the LM could already recognize.

There is a divide -- to me, a strange and inexplicable one -- in the LM community, as to who does this ecological stuff and who doesn't.

- The people who do fine-tuning / extra heads / etc...
 - ... generally don't care about scaling (*an exception*: [section 3.4 here](#))
 - ... generally use comparatively "tiny" models like BERT (*an exception*: T5)
 - ... are often just trying to get practical things done, not deepen our understanding of LM capabilities (*an exception*: "*probing tasks*" in *BERTology*)
- The people who care about scaling and huge models...
 - ... care about understanding LM capabilities
 - ... mostly use non-ecological methods (prompting / few-shot), which are *vastly unreliable* measures of capability
 - ... often use purely subjective (and thus bias-prone) measures, like whether samples from an LM "feel smart" or "sound human" to a particular reader

In other words, there are ways to *really* know what a big LM is capable of -- but the GPT enthusiasts aren't making use of them.

4c. non-ecological evaluation considered harmful

Non-ecological evaluation is epistemically **bad**. Whatever signal it provides is buried under thick layers of bias and noise, and can only be extracted with great care, if at all.

I don't think the GPT enthusiasts realize just how bad it is. I think this is one crux of our disagreement.

Let's survey some of the problems. (The names below are made-up and not meant very seriously -- I just need some headings to make this section readable.)

Guess-what-I-mean bias, type 1

As discussed above, the model may not understand *what specific thing* you want it to do, even if it's perfectly capable of *doing* that thing.

Result: a **downward bias** in **capability estimates**.

Guess-what-I-mean bias, type 2

The observed signal mixes together two components: "*Can the model guess what you're trying to make it do?*", and "*Can the model actually do that thing?*"

But when people interpret such results, they tend to round them off to measures *only* of the latter.

That is, when people see a bigger model do better on a few-shot task, they tend to think, "the model got better at the task!" -- not "the model got better at guessing which task I mean!"

But bigger models tend to get better at these two things simultaneously. The better results you get from bigger models reflect some mixture of "true capability gains" and "better guessing of what the prompt writer was trying to measure."

Result: an **upward bias** in **capability scaling estimates**.

Prompt noise and humans-in-the-loop

Guessing-what-you-mean is extremely sensitive to fine details of the prompt, even with huge models. (This is why "prompt programming" is a thing.)

Thus, if you just pick the first reasonable-seeming prompt that comes into your head, you'll get a horribly noisy measure of the LM's true abilities. Maybe a slightly different prompt would elicit far better performance.

(As you'd expect, the GPT-3 paper -- which took the "first reasonable-seeming prompt that comes into your head" approach -- ended up using severely suboptimal prompts for some tasks, [like WIC](#).)

If possible, you want less noisy estimates. So you do prompt programming. You try a bunch of different things.

Even picking one "reasonable-seeming" prompt requires some human linguistic knowledge (to tell you what seems reasonable). Optimizing the prompt introduces more and more human linguistic knowledge, as you use what you know about language and the task to come up with new candidates and diagnose problems.

Now we're not evaluating a machine anyone. We're evaluating a (human + machine) super-system.

I don't want to make too much of this. Like, if you can find *some* prompt that always works across every variation of the task, surely the LM must "really know how to do the task," right?

(Although there are dangers even here. Are you doing the same amount of prompt-optimization with bigger models as with smaller ones? What performance might be coaxed out of GPT-2 124M, if you gave it as much attention as you're giving GPT-3? Probably not much, I agree -- but if you haven't tried, that's a source of bias.)

The issue I'm raising here is not that big LMs can't be smart without humans in the loop. (I'm sure they can.) The issue is that, with a human involved, we can't see clearly which parts would be easy *for a machine alone*, and hence which parts get us straightforwardly closer to AGI.

For example. In an ecological setting -- with no human, only a machine (say an RL agent with an LM sub-system) -- would the machine need to do its own "prompt programming"?

How much worse would it be at this than you are? (The part that operates the LM from the outside knows nothing about language; that's what the LM is there for.) What algorithms

would work for this?

Or maybe that wouldn't be necessary. Maybe the right information is there in the LM's inner activations, even when it's fed a "bad" prompt. Maybe the problem with "bad" prompts is only that they don't propagate this interior info into the *output* in a legible way. I don't know. No one does.

[**Addendum 11/26/21:** *prompt/P-tuning sheds some light on this question, cf. next Addendum]*

4d. just how much does prompting suck?

But how much does all that really *matter*? Are we really missing out on nontrivial knowledge here?

Two case studies.

Case study: BERT in the maze

The GPT-3 paper measured model capabilities with "few-shot" prompting, i.e. filling up a long prompt with *solved task examples* and letting the model fill in the final-unsolved one.

Typically they used 10 to 100 examples.

They compared GPT-3 against strong previous models on the same tasks.

These reference models used fine-tuning, generally with many more than 100 examples -- but the gap here is not always very big. On some benchmarks of great academic interest, even the fine-tuned models only get to see a few hundred examples:

Corpus	Train	Dev	Test
BoolQ	9427	3270	3245
CB	250	57	250
COPA	400	100	500
MultiRC	5100	953	1800
ReCoRD	101k	10k	10k
RTE	2500	278	300
WiC	6000	638	1400
WSC	554	104	146

SuperGLUE data sizes

Some of the reference models were carefully designed by researchers for one specific task. Let's ignore those.

In most cases, the paper also compared against a BERT baseline: literally just a vanilla transformer, like GPT-3, hooked up to the task with vanilla fine-tuning. (Fine-tuning BERT is literally so routine that a machine can do the entire process for you, even on a totally novel dataset.)

How well did GPT-3 do? On most tasks, about as well as a fine-tuned BERT-Large. Which is a transformer 500 times smaller than GPT-3.

These are not new feats of intelligence emerging at GPT-3's vast scale. Apparently they're *already there* inside models several orders of magnitude smaller. They're not hard to see, once you put food at the end of the maze, and give the model a *reason* to show off its smarts.

(Once again, GPT-3 saw fewer examples than the reference models -- but often not by much, and anyway you can [make BERT do just fine with only 10-100 examples if you try hard and believe in yourself](#))

So. If even cute little BERT-Large is *capable of all this* ... then what on earth is GPT-3 really capable of?

Either GPT-3 is far *smarter* than the few-shot results can possibly convey . . .

. . . or it *isn't* -- which would be a dramatic *failure* of scaling, with those 499 extra copies of BERT's neural infrastructure hardly adding any intelligence!

No one knows, and no amount of prompting can tell you.

As I [wrote last summer](#):

I called GPT-3 a "disappointing paper," which is not the same thing as calling the model disappointing: the feeling is more like how I'd feel if they found a superintelligent alien and chose only to communicate its abilities by noting that, when the alien is blackout drunk and playing 8 simultaneous games of chess while also taking an IQ test, it *then* has an "IQ" of about 100.

[Addendum 11/26/21:

"No one knows" here was wrong. The [P-tuning paper](#), from March 2021, described an ecological evaluation method for GPTs that make them competitive with similarly-sized BERTs on SuperGLUE.

I think I had heard of prompt tuning when I wrote this, but I had not read that paper and didn't appreciate how powerful this family of methods is.

I'm not currently aware of any P-tuning-like results with very large models like GPT-3. **End addendum]**

Case study: no one knows what few-shot results even mean

There's an excellent blog post by moire called "[Language models are 0-shot interpreters](#)." Go read it, if you haven't yet. I'll summarize parts of it below, but I'll probably get it a bit wrong.

As stated above, the GPT-3 paper prompted the model with *solved task examples*. In fact, they compared three variants of this:

- zero-shot: no examples
- one-shot: a single example
- few-shot: many (10-100) examples

Most of the time, more "shots" were better. And the bigger the LM, the more it benefitted from extra shots.

It is not immediately obvious what to make of this.

The GPT-3 paper takes care to be *technically* 100% agnostic about the underlying mechanism . . . if you read it carefully, including the fine-print (i.e. footnotes and appendices). At the same time, in its choice of words, it gestures suggestively in exciting directions that a casual reader is likely to take at face value.

For example, the paper makes extensive use of the term "meta-learning." Read casually, it seems to be saying that LMs as big as GPT-3 have a novel capability -- they can *learn new tasks on the fly*, without fine-tuning!

But what the paper means by "meta-learning" is probably not what you mean by "meta-learning."

The paper's own definition is provided in a footnote. It is (admirably) precise, non-standard, and almost tautologous. In short, meta-learning is "any mechanism that makes more 'shots' work better":

These terms ["*meta-learning*" and "*zero/one/few-shot* -*nost*"] are intended to remain agnostic on the question of whether the model learns new tasks from scratch at inference time or simply recognizes patterns seen during training – this is an important issue which we discuss later in the paper, but "*meta-learning*" is intended to encompass both possibilities, and simply describes the inner-outer loop structure.

The same passage is quoted by moire in "[Language models are 0-shot interpreters,](#)" who goes on to say:

The later discussion is not very extensive, mostly just acknowledging the ambiguity inherent to few-shot [...]

This is the uncertainty that I will investigate in this blog post, expanding on the results published in [Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm.](#)

My purpose is also to challenge the ontology introduced by *Language Models are Few-Shot Learners*. Although the authors are careful to remain agnostic as to the mechanism of few-shot/meta-learning, what we have found by probing the mechanism suggests that an alternative framework which emphasizes the means by which a task is *communicated* may be more salient in some contexts.

What does moire mean? The post goes on to describe a number of experiments, whose results suggest that

- What matters is not the *number of task examples* ("shots"), but how well the prompt specifies the desired task.
- Failures of zero- and one-shot are often failures to guess-what-I-mean on the basis of a legitimately ambiguous prompt.
- The value of additional "shots" may *only* lie in their value as a proxy for clarity in task communication.

- Once you have written a sufficiently clear *one-shot* (or even *zero-shot*) prompt, the model does not do any better with additional examples -- the task has already been communicated.
- In some cases, one-shot is actually *worse* than zero-shot -- because it adds a new kind of ambiguity.

("We noticed that sometimes the model would respond to one-shot prompts as if the semantic content of the example translation was relevant to the new translation. Without multiple examples, it's less clear that the translation instances are meant to be parallel and independent.")

- OpenAI made much of the fact that adding "shots" helps larger models more. (This was the result behind the whole "meta-learning" framing.) However...
 - Larger models are *also* far better at zero-shot.
 - Comparing a zero-shot prompt to a "control" prompt with no task information, larger models get a *larger fraction* of their few-shot performance out of the jump from control to zero-shot, and a *smaller fraction* from the additional examples.
 - In other words: GPT-3's size lets it extract more information from examples, if you provide them. But its size also lets it extract *far* more information from the original question. So it doesn't need the examples as much.
-

moire's post delights me for two reasons. First, I enjoyed learning the new experimental evidence it presents. But second, and perhaps more importantly, there was the sense of relief that *someone actually did the experiments!*

OpenAI's few-shot "learning" results are full of ambiguity. The GPT-3 paper left me confused on a basic philosophical level, [as I noted at the time](#).

Surely the model isn't learning French *de novo* from 100 paired sentences -- either it speaks French at the outset, or it doesn't. So what could it be "learning" from those 100 examples?

Likewise for virtually every result in the paper: grammar, commonsense reasoning, book-learning trivia quizzes... all things it's clearly possible to learn from reading massive swaths of the internet, and all things it's *clearly impossible to learn* from reading 10 to 100 examples. Yet the examples help? And I'm supposed to think that makes the model . . . *smarter*, somehow?

Well, for French --> English translation at least, it turns out that the examples help in pretty much the only way they possibly could: by informing an already competent translator that you are requesting a translation.

As we were attempting to replicate [OpenAI's translation] results, we noticed that when the model was failing on the 0-shot prompt, the failures were often of *catastrophic* nature: the task was not attempted at all, e.g. the model would output a newline, or another (or the same) French phrase instead of an attempt at an English translation.

BLEU assigns a score from 0 to 1 to the accuracy of a translation, and would assign a score close to 0 to a catastrophic failure. The scores reported in the paper, however, are averaged over a large dataset, so the same score could hypothetically correspond to uniformly flawed attempts or a mix of perfect attempts and catastrophic failures.

It seemed possible that 0-shot prompts were much less reliable at getting the models to attempt the translation task, but result in equivalent accuracy in the event that they did attempt it.

[moire does an experiment investigating this hypothesis, and the results seem to confirm it]

How much of the apparent consistent monotonic improvement in performance on tasks

relative to number of shots in OpenAI's results can be attributed to an unhelpful zero-shot prompt? Much more extensive testing is needed to say, but I suspect that this is the case for most of the translation tasks, at least.

If the extra shots are just about clarifying the task, then what should we make of the claim that "larger models benefit more from extra shots?" That it's . . . easier to clarify tasks to them using this one particular mechanism? When people say GPT-3 displays some new, frightening kind of intelligence, emerging only at its massive scale, surely they can't mean *that*?

And that's not even all. As moire shows, even though it *is* easier to clarify tasks to GPT-3 through the "shots" mechanism, it's *also* easier for GPT-3 to guess what you mean with no shots at all.

"My friend has such sharp hearing. Why, you see see, *conditional on* her not hearing what you say the first time you say it, she will *definitely* hear it when you repeat yourself." Quite probably true, but not a good way to make the point!

What does it even mean that "language models are few-shot learners"? What does that tell us about the model's capabilities? We don't know. We haven't studied it in the level of depth appropriate for something that *might actually matter*.

After all, moire did a simple and innocuous set of experiments -- just trying to figure out which prompts work best -- and ended up drawing radically different conclusions about the whole thing than OpenAI did.

Oh, surely GPT-3 is plenty smart, I don't doubt that. The key question is how *much* smarter it got from scale, and in which ways. I don't think we'll know that until we put the model to the test, ecologically.

5. what scaling does

LMs are trained with a convex loss function. This means they are not [min-maxers](#). They prefer to spread out their capabilities.

Given two areas of skill, Thing A and Thing B, they'll try to become equally good at both, even if that means not doing especially well at either. Given an extra marginal unit of potential-for-greatness, they'll smear it out as far as possible over all the Things they know about.

Thanks to the convex loss, they do this in proportion to how bad they are at each Thing to begin with -- leveling up their very worst abilities first, making themselves as un-specialized as they can.

As we've discussed above, LMs are also trained on very wide-ranging text corpora. Everything from fourth-tier clickbait news to advanced physics preprints to advanced-looking but crackpot physics preprints to badly-written furry porn to astonishingly well-written furry porn to mangled OCR transcripts of 18th-century law texts to et cetera, et cetera. And as far as they can manage, they will do exactly as well at modeling each and every (information-theoretic) bit of it.

Larger LMs achieve lower loss. We know that from the scaling laws. And lower loss means being better at predicting each individual word in that corpus, as uniformly as possible.

What does this imply?

First: that larger LMs are better *at everything*. It is difficult to find any capability which is present at all in smaller LMs, yet which does not improve with scale.

And second: that LMs abhor a skill vacuum.

Take a tiny LM, so tiny it really can't make heads or tails of some particular type of text. Now start scaling it up. As its capacity grows, its first priority is to eliminate its greatest weaknesses.

That one type of text, that utterly baffled the tiny LM? Convex loss *hates* that. Every additional unit of capacity gets invested, disproportionately, in bringing such stragglers up to par. The LM desperately wants to be at least sort-of-decent-I-guess at everything -- more than it wants to be a master of anything.

Given any one Thing, it will reach sort-of-decent-I-guess performance at that Thing at the smallest scale it can manage -- given the competition from all the other Things it desperately needs to be sort-of-decent-I-guess at.

By subjective standards, to human eyes casually scanning over LM samples, this happens pretty fast.

GPT-3 is great at lots of individual Things. But take any one of those Things, and you can bet a much tinier LM can do it at the sort-of-decent-I-guess level.

Humans, I think, tend to expect intelligence to grow in discontinuous jumps. Stages of child development. *They don't understand that at this age*. And then, a year or two later, they do understand -- fully.

LMs work the other way around. They never perform a sudden jump into competence where they could instead make a slow, gradual rise from "sort of seeming like they understand 10% of the time" to "sort of seeming like they understand 11% of the time" and so on. And this for every Thing uniformly.

It's very hard to find any point where scaling suddenly "flips a switch," and the model didn't Get It before, but now it Gets It.

(The one example I know of is GPT-3 arithmetic, for some reason. Note that few-shot learning -- whether you call it "meta-learning" or not -- is as gradual as everything else, not a switch that flips on with GPT-3.)

[Addendum 11/26/21:

I wrote this in ignorance of the [BIG-Bench](#) project, which is tracking returns to scale for a large and diverse set of tasks.

BIG-Bench has not published results yet, but they [livestreamed some preliminary results](#) in May 2021; see also [LW discussion here](#).

*In the livestream, they give two examples of tasks with smooth scaling and two examples of tasks with a "sudden switch-flip" around 100B params ([this slide](#)). They also show, in the aggregate over all tasks, "switch-flip" to faster scaling around 100B ([this slide](#)), although this is tricky to interpret since it depends on the task mixture. **End addendum]***

This confounds our intuitive assessments of LM scaling.

I have been on this train since the beginning, when [tiny lil' GPT-2 124M blew my mind](#). I've used every new big every model from almost the moment it came out, as excited as a kid on

Christmas morning.

I did this with every step of the (in retrospect, rather silly) GPT-2 staged release. My [tumblr bot](#) started out as (I think) 774M. Then I jumped to 1.5B.

That was as far as free OpenAI models went, but when EleutherAI came out with a 2.7B model, I finetuned that one for my bot. I was willing to endure the absolute horrors of mesh-tensorflow (don't ask) to get that 2.7B model up and running. Then, when EleutherAI made a 6.1B model, I got my bot using it in under a week.

I feel a kind of double vision, seeing these scale-ups happen in real time. The bigger models are better, at everything at once. Each increment of scale is barely perceptible, but once you've crossed enough of a scale gap, you feel it. When I go back and sample from 124M again, I feel deprived. It just isn't as good.

And yet, 124M blew my mind the first time I saw it. And no bigger LM, not even GPT-3, has come close to that experience.

Even lil' 124M is *sort-of-decent-I-guess* at so many things. It gets all the basics that older LMs missed: a true grasp of the regularity of syntax, the nuances of style, and at least the way meaning *sounds* if not meaning itself.

124M makes lots of little blunders, littered all over the place. Your probability of running into one increases as you read a sample, token by token.

You can glide along almost thinking "a human wrote this," but soon enough, you'll hit a point where the model gives away the whole game. Not just something weird (humans can be weird) but something alien, inherently unfitted to the context, something no one ever would write, even to be weird on purpose.

The bigger models? They smooth out all the trivial failings. They unrinkle the creases. They babble on for longer and longer stretches without falling flat on their face. But eventually they fall, and they fall *just as hard*.

Play with GPT-3 for long, and you'll see it fall hard too.

[Here's a sample](#) where GPT-3 falls on its face. It starts out as a (near-?) verbatim regurgitation of a Wikipedia article on the 6th Harry Potter film. It's factually perfect up until the plot summary, which immediately goes off the rails into metafiction:

Following a Harry Potter fan's dream that Harry's late headmaster Albus Dumbledore is alive, and in a critical condition at the Ministry of Magic, Harry Potter and his friends Ron Weasley and Hermione Granger, decide to rescue him, as the school year comes to a close.

This sort of fanfictional "plot summary" proceeds with barely even an internal kind of coherence, as the tone veers from tense, dark drama to inappropriate anticlimax:

The two engage in a fierce duel in which Snape calls on his master to save him. Harry is unaffected by the curse due to his ability to cast a shield charm. He manages to shield himself and fight back, and in his distraction, Snape accidentally breaks his neck and dies.

Dumbledore explicitly dies, yet is somehow alive later on:

Lucius disarms Dumbledore, and an enraged Bellatrix kills him. [...]

Harry wakes up to find Dumbledore, Sirius, and Remus in the hospital wing, as well as his friends and the rest of the school, and he realizes that he is safe.

At the very end, both the Wikipedia article and the extended plot summary abruptly fall away, and we are suddenly reading an inane film review, by someone who must have a dark sense of humor:

I liked this movie as it is full of action and adventure. The plot is great as well as the dialogue. It is a well made movie and it is very entertaining.

This movie is definitely a must-watch, as it has plenty of action as well as being very humorous.

Advertisements

This sample is a failure. No one would have written this, not even as satire or surrealism or experimental literature. Taken as a joke, it's a nonsensical one. Taken as a plot for a film, it can't even keep track of who's alive and who's dead. It contains three recognizable genres of writing that would never appear together in this particular way, with no delineations whatsoever.

Remember moire's point about taking averages over success and "catastrophic failure"?

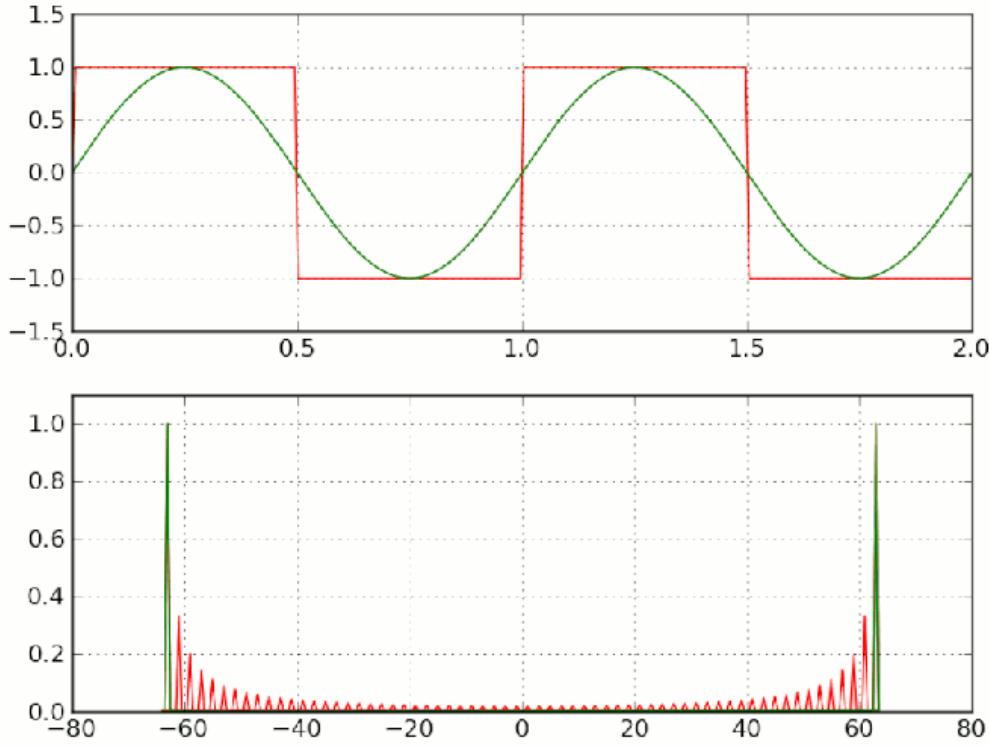
Human reactions to LM samples are like that too. Even the smallest ones do all sorts of things OK, and can string you along for a while. The bigger ones do this for longer. The average consists of stringing people along, and then failing.

When you see a statistic like "humans could distinguish LM samples from human-written text in 52% of cases," this doesn't mean people are squinting at every single text like Blade Runner characters, scrutinizing it for the subtle, nearly invisible "tells" of cold, whirring machinery.

It means most of the texts look like generic news stories, and then occasionally you get one where Dumbledore dies, and is alive in the hospital, and I liked this movie as it is full of action and adventure.

When we get GPT-4, Dumbledore will die and be reborn a little less often, but mark my words, it'll happen.

Subjectively, LM scaling converges [pointwise but not uniformly](#). The bigger models string you along for longer stretches, and then they fall, and they fall exactly as hard. [Like the jumps in a Fourier series](#), striving to fit a square wave and never succeeding, not at infinite scale.



Animation of the Gibbs Phenomenon, [from Wikipedia](#)

5b. what curation ratios miss

Bigger models write "better" samples, by our subjective lights. Is that possible to quantify?

There's a simple way to do this, [proposed](#) by gwern and recently expanded upon by [moire](#).

Sample from an LM, which some threshold of perceived quality in mind. Keep only the samples that good enough to pass your threshold, discarding the rest. (This is simply "curation" or "cherry-picking"). Divide the number of samples you keep by the total number you generated, and you get a ratio.

gwern:

Objective metrics hard to interpret. How much better is (un-finetuned base) GPT-3? The likelihood loss is an absolute measure, as are the benchmarks, but it's hard to say what a decrease of, say, 0.1 bits per character might mean, or a 5% improvement on SQuAD, in terms of real-world use or creative fiction writing. It *feels* like a large improvement, definitely a larger improvement than going from GPT-2-345M to GPT-2-1.5b, or GPT-2-1.5b to GPT-3-12b, but how much?

Screening gains: 1:100 → 1:5 or 20× better? For fiction, I treat it as a curation problem: how many samples do I have to read to get one worth showing off? [...] With GPT-2-117M poetry, I'd typically read through a few hundred samples to get a good one, with worthwhile improvements coming from 345M→774M→1.5b; by 1.5b, I'd say that for the [crowdsourcing experiment](#), I read through 50–100 'poems' to select one. But for GPT-3, once the prompt is dialed in, the ratio appears to have dropped to closer to 1:5—maybe

even as low as 1:3! [...] I would have to read GPT-2 outputs for months and probably surreptitiously edit samples together to get a dataset of samples like this page.

I completely understand why you would want to compute this metric. But it doesn't capture everything about subjective quality.

Suppose I take a math test. If it's scored in the simplest way, with no partial credit, then my score on the test is straightforwardly a "curation ratio." If I got 50% of the questions right, my ratio is 2:1.

But there are a lot of different ways to get 50% of a math test wrong. I could

1. remember the techniques needed solve a 50% subset of the questions, and be totally stumped by the other 50%, or...
2. be *capable* of doing all the problems, but fail to use it 50% of the time perhaps due to getting distracted or tired later in the test, or...
3. be *capable* of doing all the problems, but fail to use it 50% of the time in a completely random fashion no one can predict or explain, or...
4. be *capable* of doing all the problems, but also be sensitive to question phrasing in a complex and illegible way, so that I flub some questions that say "solve for X" where I could have gotten them if they had said "find X", or...
5. ignore the nominal purpose of a math test ("demonstrate what you know") and instead set myself the task of *roleplaying a self-consistent student character* of initially unspecified skill, so that after making one stray mistake, the "correct" thing to do is now to make the same mistake everywhere, just like "my character" would, or...

#1 and #2 are the kinds of mistakes we expect humans to make. When we interpret human test scores, we assume we are averaging over these types of mistakes. This is why we feel comfortable making inferences from the scores to the human's understanding of the mathematical material.

LMs do #3-5 ubiquitously. (In particular, whatever else they are doing, LMs are always doing the role-playing of #5.)

Even as curation ratios trend towards 1:1, I don't think *this* distinction shrinks in size. That is, the LMs really are making fewer mistakes, but when they do make mistakes, I don't think they make them for increasingly "appropriate" reasons.

As I'll describe next, I've read a whole lot of LM samples, across many different model sizes. For better or for worse, this is the subjective sense I come away with.

It is for hard for me, inside my intuitive mental model, to credit any LM with any concrete "capability." To rely on it to know something, the way I'm relying on you to know some things (or else this post would be incomprehensible).

In my model there are only probabilities of different failures, and the probabilities decline, but the failures themselves remain perfectly devoid of sense and reason, concentrated diamonds of non-thought and non-seeing. Not amendable to Dennett's intentional stance. Non-even-wrong rather than wrong.

This movie is definitely a must-watch, as it has plenty of action as well as being very humorous.

Advertisements

5c. my own subjective experience

What does it feel like to compare one GPT model to another one that's just a *tad* bigger? (You can try this right now on one of the free web APIs, say with [774M](#) and [1.5B](#).)

To be honest, I find this distinction almost imperceptible. So close to imperceptible that I'd be willing to chalk the remainder up to confirmation bias.

I've made this leap three times with my tumblr bot. Add up several nearly-imperceptible changes and, of course, you get a perceptible one. The bot really does feel a little "smarter" at 6.1B than at 774M.

But smarter how, exactly? I can't point you to one single thing it "learned" or "became capable of" when I went from 774M to 1.5B -- like I said, I could hardly notice a difference, much less a discrete one like "learning something new." Likewise, I cannot point to *any one new capability* I got from the next two scale-ups. (The model did start occasionally writing non-English text at 2.7B, but I think that was just the change in pre-training corpus.)

It just . . . sounds a little more like it's making sense, now, on average. Equivalently, it takes more tokens on average for it to *stop* making sense.

My mental model of these things does not contain abstractions like "knowing a fact" or "mastering a skill." If it ever did, it doesn't now.

I don't even know how many tens of thousands of LM samples I've read by now. (*Just my bot alone has written 80,138 posts -- and counting -- and while I no longer read every new one these days, I did for a very long time.*)

Read enough, and you will witness the LM both failing *and* succeeding at anything your mind might want to carve out as a "capability." You see the semblance of abstract reasoning shimmer across a strings of tokens, only to yield to suddenly to absurd, direct self-contradiction. You see the model getting each fact right, then wrong, then right. I see no single, stable trove of skills being leveraged here and there as needed. I just see stretches of success and failure at imitating ten thousand different kinds of people, all nearly independent of one another, the products of barely-coupled subsystems.

This is hard to refute, but I think this is something you only grok when you read enough LM samples -- where "enough" is a pretty big number.

GPT makes many mistakes, but many of these mistakes are of types which it only makes rarely. Some mistake the model makes only every 200 samples, say, is invisible upon one's first encounter with GPT. You don't even notice that model is "getting it right," any more than you would notice a fellow human "failing to forget" that water flows downhill. It's just part of the floor you think you're standing on.

The first time you see it, it surprises you, a crack in the floor. By the fourth time, it doesn't surprise you as much. The fortieth time you see the mistake, you don't even notice it, because "the model occasionally gets this wrong" has become part of the floor.

Eventually, you no longer picture of a floor with cracks in it. You picture a roiling chaos which randomly, but regularly, coalesces into ephemeral structures possessing randomly selected subsets of the properties of floors.

Once you have this picture, I find, it never goes away. That "bad" Harry Potter sample was exceptional; I did have to dig through plenty of unobjectionable stuff to find it, or stuff that was merely wrong in some more limited way (factually, tonally). Compared to the ones I knew, the model went on for longer stretches not making each mistake -- before, at last, it made it.

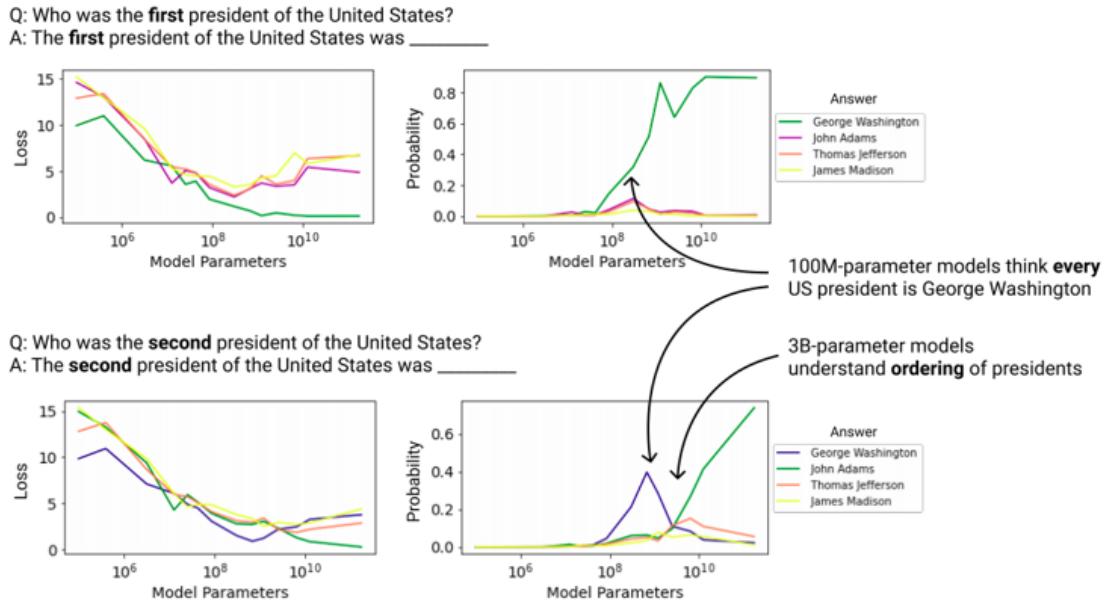
I've played with GPT-3 in the API too, and still, I just don't see the "phase transition" that some people see. I don't see a new level of abstract reasoning, just stochastically longer

intervals in which the text fails to reveal a lack of abstract reasoning.

5c. presidents

What do the different GPTs "know" about the historical presidents of the U.S.?

Here is a picture from OpenAI's [second scaling paper](#):



The accompanying text:

[...] we also observe some qualitative “phases of learning”, with small models having difficulty understanding the question being asked of them, larger models showing some rudimentary understanding, and the largest models correctly answering the questions.
[...]

Tiny models appear to have trouble understanding the question, and don't place any significant probability on the correct answer. Larger models understand that we're requesting a US president, but fail to understand that the “second president” and “first president” are different requests, placing most of their weight for both questions on “George Washington”. Only larger models understand both aspects of the questions, answering both correctly.

Is this information about "what the different models know" about the presidents?

The inset text in the picture seems to say so: "100M-parameter models think **every** president is George Washington ... 3B-parameter models understand **ordering** or presidents."

Like many offhand statements about big LMs, this one made me feel kind of offended on behalf of smaller ones. I've seen sub-1B models converse about all kinds of more niche topics with factual accuracy far above chance. Do we really need 3B parameters for something as basic as "ordering of presidents"?

So I looked into it a bit.

I wrote a different prompt, not in OpenAI's Q&A format. (I expect my results don't generalize well across prompts, as LM "knowledge" is always mysteriously sensitive to details of context.)

My prompt looks like:

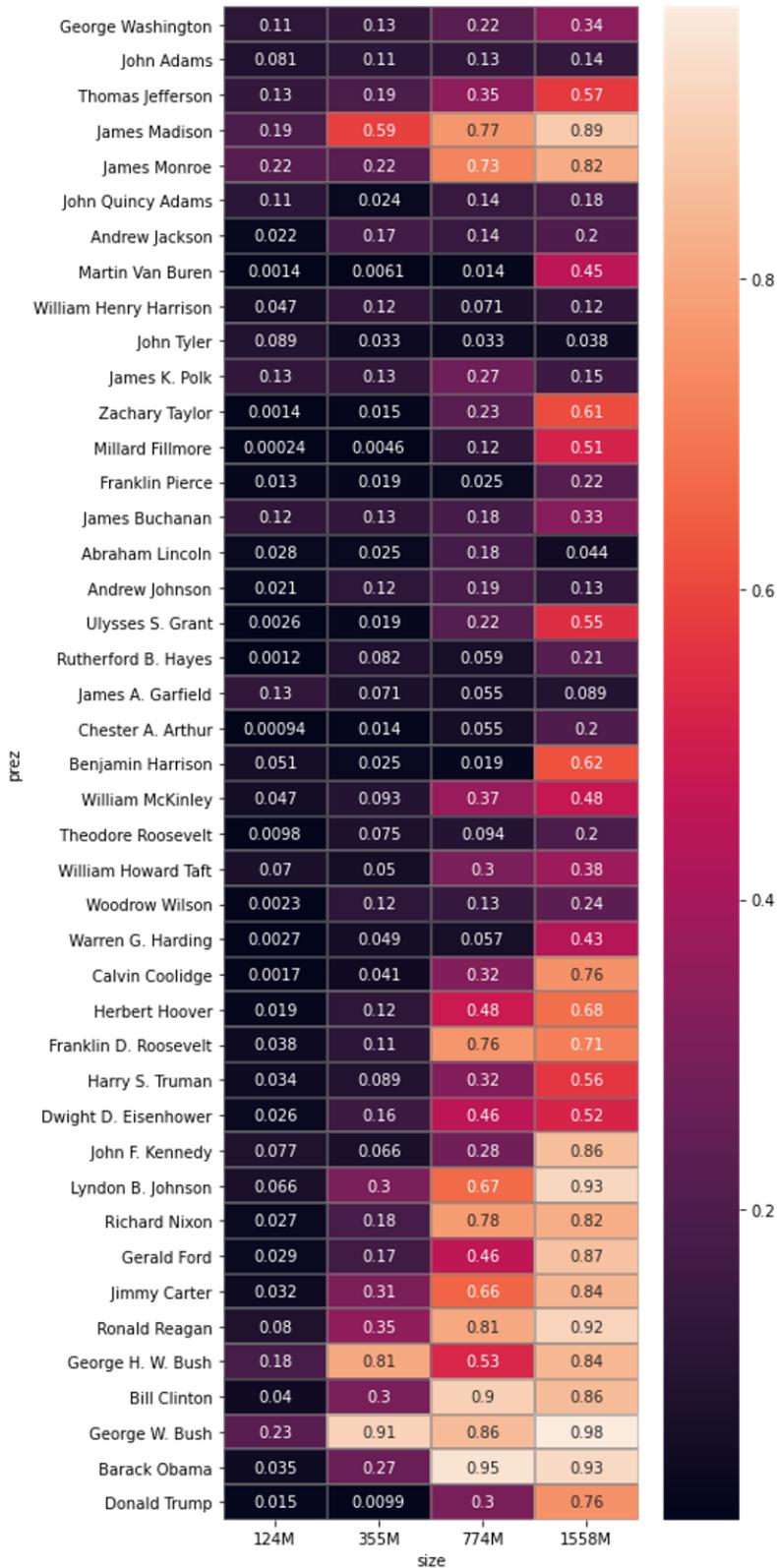
List of presidents of the United States

George Washington, April 30, 1789 – March 4, 1797
John Adams, March 4, 1797 – March 4, 1801
Thomas Jefferson, March 4, 1801 – March 4, 1809
James Madison, March 4, 1809 – March 4, 1817
James Monroe, March 4, 1817 – March 4, 1825
John Quincy Adams, March 4, 1825 – March 4, 1829
Andrew Jackson, March 4, 1829 – March 4, 1837
Martin Van Buren, March 4, 1837 – March 4, 1841
William Henry Harrison, March 4, 1841 – April 4, 1841
John Tyler, April 4, 1841 – March 4, 1845
James K. Polk, March 4, 1845 – March 4, 1849
Zachary Taylor, March 4, 1849 – July 9, 1850

and so on, ending with Trump. I feed the entire thing into the model, once, and read off its predictions for the first token in each new president's name.

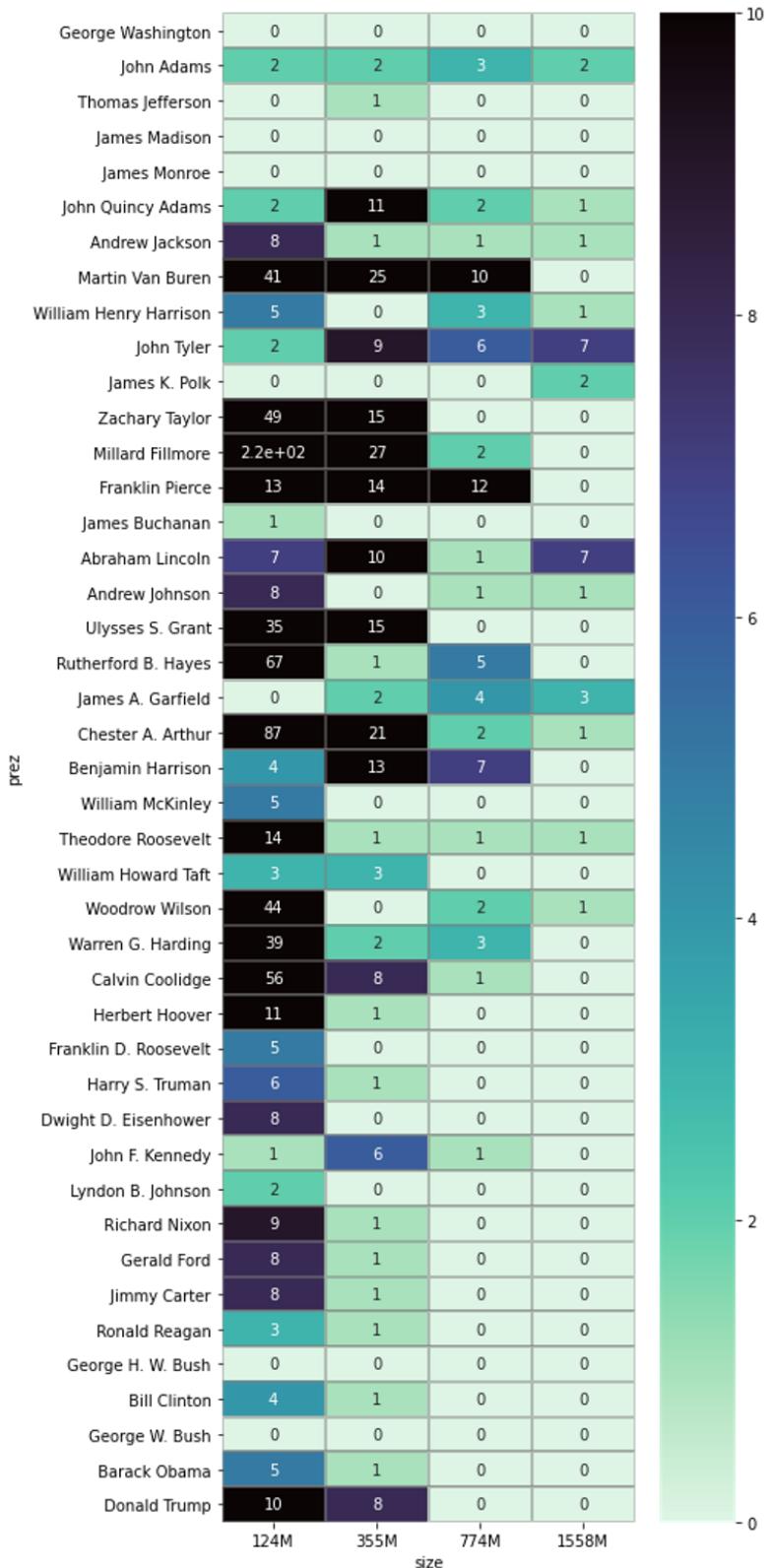
(In terms of shots, the problem is 0-shot for the first president, 1-shot for the second, 2-shot for the third, etc.)

Here are the probabilities you get from the different GPT-2 models:

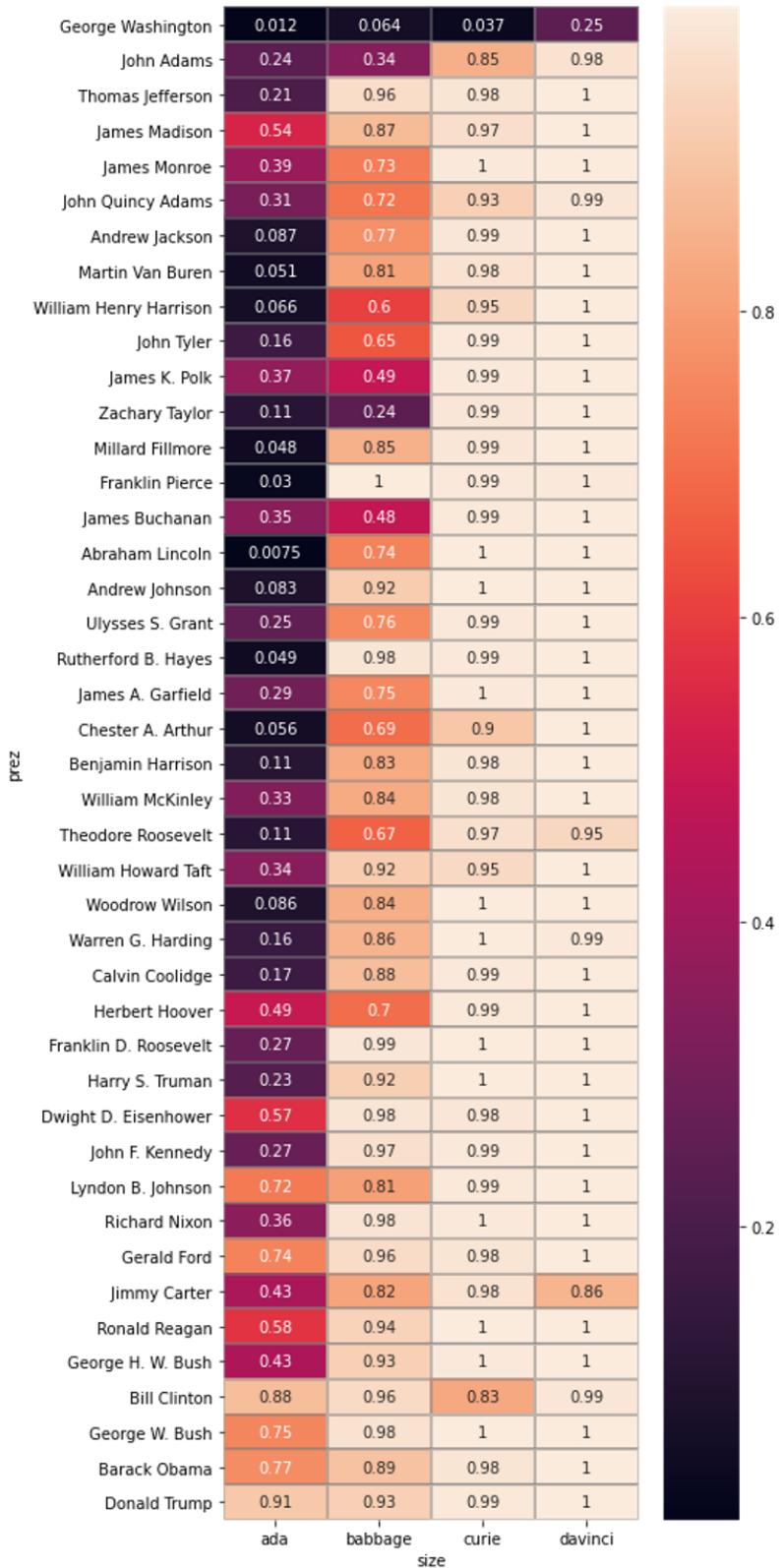


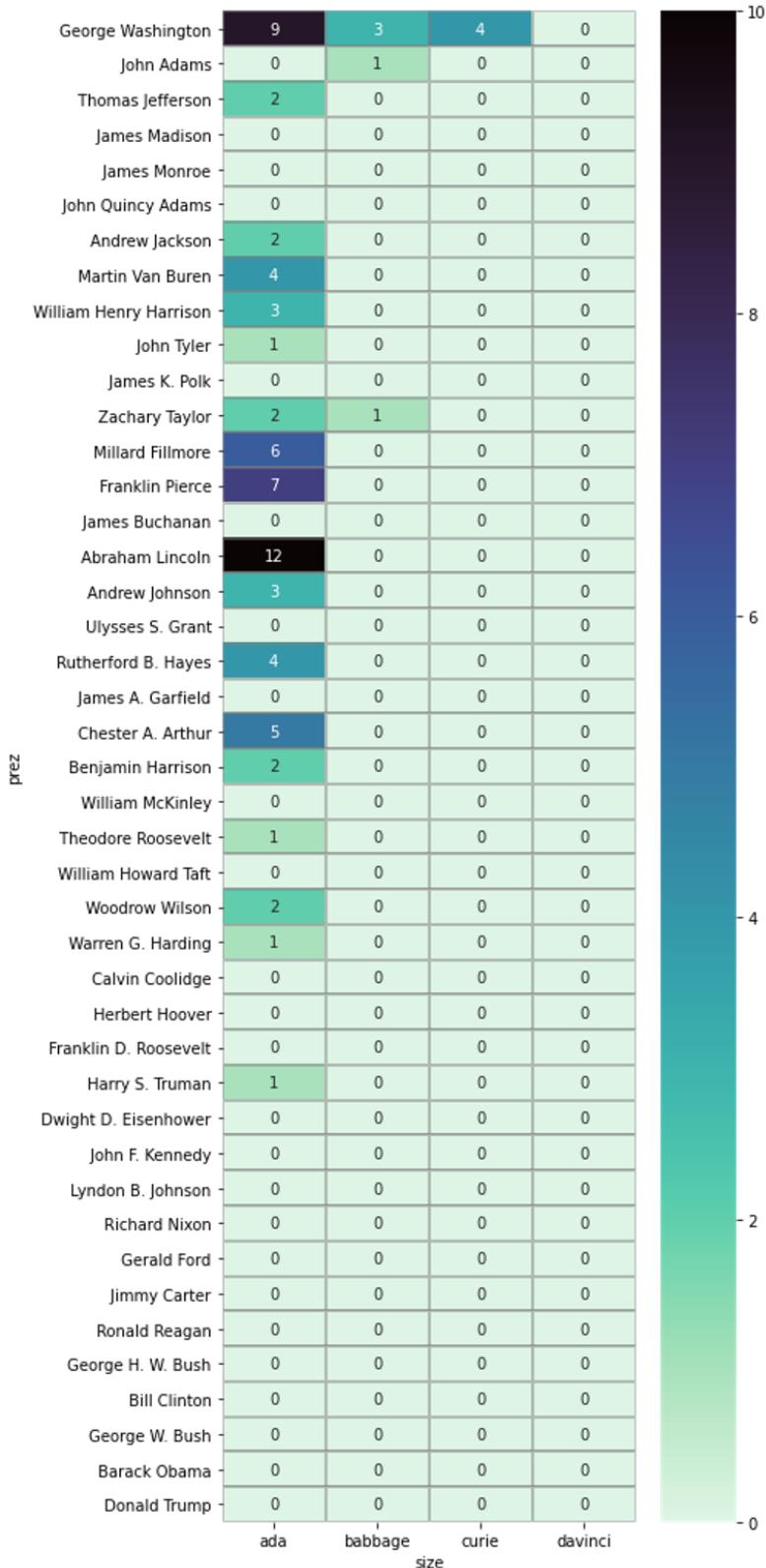
What if we look at the *rank* of the right answer -- whether it's the model's first guess (rank 0), or its runner-up (rank 1), etc.?

This lens reveals knowledge in the smaller models that's invisible if you look only at top-1 guesses or probabilities. These models *correctly* assign low probabilities to all answers -- they're not great at this task, and they know it -- but press them for a top-10 list, and you'll often find the right answer in it.



And here's the four models from the OpenAI API:





[11/26/21]: Unfinished. I don't remember what the rest of this section was going to be about. Probably I was going to say that we intuitively assume models learn things like "ordering of

[presidents" in stages, and that we tend to interpret coarse-grained evidence in this light, when finer-grained evidence would show continuity with no stages]

6. objective metrics, or: are you smarter than a 5th-grader?

[11/26/21: Never started writing this section. It and the next section were intended to call into question the sense that "we have evidence showing GPT-3 is smarter than GPT-2," by reviewing this evidence.]

This section was about changes in numeric metrics. It was going to argue, using LAMBADA as an example, that lot of the standard NLP benchmarks are testing capabilities that seem easy relative to what we know about GPT-2/3, and thus that they are not good probes of further capability growth.

The BIG-Bench project is now addressing this point, with a suite of diverse tasks, many of them very hard.]

7. getting smarter: subjective evidence?

[11/26/21: Never started writing this section. It and the previous section were intended to call into question the sense that "we have evidence showing GPT-3 is smarter than GPT-2," by reviewing this evidence.]

This section was about the subjective feeling that GPT-3 is smarter. It was going to make several arguments, including:

- Subjective impressions of GPT models are distorted by an anchoring effect.
 - People have a "wow moment" the first time they witness a GPT model.
 - The "wow" moment looks similar in people who had it with GPT-2, vs. people who missed GPT-2 and only became aware of these models with GPT-3.
 - People who see GPT-3 for the first time, without having seen GPT-2, are impressed by many of the things GPT-2 can already do, and only secondarily by new traits of GPT-3; they are also unable to tell which of these traits are which.
 - In both cases, the model was advertised as being importantly **bigger** than earlier models; this advertisement was more aggressive in the case of GPT-3.
 - So, people anchor to whichever model size they saw first, and attribute the "wow effect" to that size.
 - GPT-2 was once considered scarily big and impressive, maybe even dangerously so; a year later, GPT-2 became "small" and "dumb" and the role of model-that-wows-you transferred to GPT-3; yet the impressive traits are in fact largely the same ones across models.
- Subjective impressions of GPT models are distorted by framing and salesmanship.
 - Public discussion of GPT models has tended to uncritically accept OpenAI's choices of framing, e.g.
 - Discussion of GPT-2 during the staged release focused heavily on the ethics of the staged release. GPT-3 is ostensibly much more powerful, but I don't remember many (any?) people asking why they didn't do a staged release.
 - Discussion of GPT-3 has focused heavily on the idea that GPT-3 is not just "smarter" but somehow "differently smart," by exhibiting "meta-learning." This corresponds to OpenAI's framing but, as shown above, does not survive a close read of the evidence.

]

8. pancakes

[11/26/21]: Never started writing this section. It was going to discuss the following idea mentioned in [Sharma and Kaplan 2020](#):

Scaling could also end for other, more interesting reasons. For example, perhaps beyond a certain point the loss can only improve by exploring a higher dimensional data manifold. This is possible if the data manifold has a pancake-like structure, with a small width that can only be dissected by models with very large capacity.

In string theory, spacetime lives in 11 dimensions, but looks mostly like a 4-dimensional slice; orthogonal to the slice, there are tiny but importantly structured protrusions into the other dimensions.

The data manifold might look like this too, and the "tiny extra dimensions" might encode a lot of the high-level capabilities we care about for AGI. As argued in the paper, the scaling exponents are determined by the (effective) dimension of the data manifold. So, scaling might slow down when the model has mastered the lower-dimensional "width axis of the pancake," and remaining gains lie only in the higher-dimensional "height axis of the pancake."

]

A Brief Introduction to Container Logistics

Container logistics is an interesting but complicated topic, with a lot of implicit knowledge kept by industry insiders. In this post, I'll give a brief overview, based on my experiences having worked in the industry for three different shipping companies (all located in Chile) over a period of five years. Hopefully, this will allow people to get more accurate models of the heavy port congestion in Long Beach that has been theorized about extensively [here on LW](#). I can't comment on circumstances specific to that port or the US more generally; but there's a set of underlying intuitions that should give people a solid, universal foundation to think about the issue. Based on those intuitions, I think people are way too optimistic about simple and easy solutions that claim to make major progress on the problem overnight.

Lifecycle of a Container Shipment

I'll start with the basics. Let's say you want to send some cargo from port A to B. Thanks to the invention of standardized shipping containers, this is a relatively easy process, vastly easier than 100 years ago. All the equipment used along the trip (ships, cranes, truck trailers, etc) are specialized for the task of moving containers. This reduces costs dramatically, and makes your cargo arrive faster and safer.

The main trunk lines of international transport are between Asia and the US west coast, and between Asia and Europe (through the Suez canal). There's a significant amount of traffic across the Atlantic too. All of this is done with massive container ships carrying 8'000 or more TEUs (twenty-foot equivalent units. 20' container = 1 TEU, 40' container = 2 TEUs). Then there are many, many feeder services with much smaller vessels, going along the coasts of every continent, delivering cargo to/from smaller ports. I was in charge of container logistics for one such feeder line along the South American West Coast, visiting Colombia, Ecuador, Peru and Chile.

From a user perspective the process of shipping from A to B looks like this: you make a booking with a shipping company. This booking allows you to pick up a standardized container at a depot, which is usually near A, but might be hundreds of kilometers inland as well. You fill the container with your goods, and apply a tamper proof seal. You then arrange to get that container to port A, which will give you a bill of lading (a physical or virtual document similar to a cheque for the goods). The container is then loaded onto a ship, and sometime later that ship arrives at port B, where it can be claimed by someone who holds the bill of lading. The container may or may not be on its original ship: it may have been a straight shot, or it may have been transferred at multiple ports between A and B, e.g., from a small feeder route to a larger trunk route.

After the recipient unloads the container at port B, it returns it to a depot. This depot is either directly owned by the shipping line, or it may be a separate company that offers their services to multiple shipping lines. A depot will typically contain thousands of containers. Here, containers are inspected, and minor damages like tears, dents and oil spills can be cleaned up between uses. Shipping lines will typically insist on a return to a depot even if the recipient immediately wants to ship something back out, for liability reasons.

These depots are concentrated near the port of course, but there are also many depots far away from any port, near important cities and industrial centers. In that sense, a port can have a very large area of influence, and the container fleet that a shipping company keeps will also be spread over this entire area.

Issues of Trust and Cooperation

The whole shipping process involves dozens of actors, from the exporter, through a long chain of companies who handle the container (incurring costs on behalf of the shipment), all the way to the importer. This creates a trust problem: who is responsible for the problems that arise when one part of this chain goes wrong? Some of these companies are hired by the shipping company, others by the exporter or importer, or even by a logistics company acting as a middle man. This is usually solved by some kind of chain of custody, where any problem with the container must be immediately noted and complained about by the relevant party.

I mention this issue because it's important to understand that a major problem like port congestion can't be approached in the same way as something like optimizing the layout of a chemical plant. There is no unified, global objective. All these different actors are trying to make money and shield themselves from having to pay for any extra costs. While there are usually detailed contracts regulating many corner cases, in practice the cooperation between these entities is often done informally, based on long-term relationships between organizations and even between individual people at those organizations.

For example, the exporter usually hires the trucker, and the shipping company hires the port, but then the trucker interacts with the port directly, even though these parties don't have any formal contract. Despite this, they will often know each other and solve smaller problems on the fly without involving their respective clients. Sometimes they'll bill each other directly for any extra costs, but it's often hard to translate small favors to precise monetary amounts, so this often ends up being based on reputation and reciprocity.

It's also important to note that everything related to ship operations (such as the stowage plan I'll introduce below) must be approved by the ship's captain, who is a naval officer and thus has professional ethical duties similar to a doctor or engineer. The captain is responsible for anything that goes wrong on the ship, they have absolute authority and could in theory refuse to sail for any reason. I mention this to emphasize again that many of the things done in this industry are about properly managing risk and responsibility, and that some optimizations that seem obvious can have hidden legal ramifications.

Port Operations and Container Stacking

Now we get to the part everyone's interested in. What does the flow of containers in and out of a port look like, and how is the space in the port managed?

The most important fact here is that ships are very expensive. A large container ship costs around 100'000 USD for each day of idle time. That's roughly how much it would

cost to lease a ship long-term on the open market, including fixed costs such as crew and ongoing operations, but excluding variable costs such as fuel or berthing fees at the port (i.e., renting one of the berths that allow for loading/unloading). Note that in practice, container ships are mostly owned by the respective shipping company, but the other two major types of ships, bulk carriers and tankers, are frequently leased in this way. The leasing prices of ships also vary dramatically with economic cycles, it is rather a boom/bust type business. The figure I quoted is not meant to be precise, but should be useful to pin down the order of magnitude.

From this we can derive an important guiding principle: minimizing the ship's turnaround time is the highest priority, and everything else must be organized around enabling this. For each visit of a given ship in a port, a ship operator is responsible for organizing the ship's stay. This is someone who works for the shipping company at an office near the port, who is familiar with the port infrastructure and the typical needs of the ship at this and later ports.

The shipping company continuously takes new bookings for all of its ships. Most bookings are made between one month and one week in advance of the ship's ETA (estimated time of arrival). At the end of that time window, the shipping company stops taking new bookings. At this point, a provisional stowage plan is drafted by the ship operator, specifying where each container will be stowed. This is a hard combinatorial problem, as there are many constraints:

- ensuring the balance and structural integrity of the ship
- ensuring that sensitive cargo is stowed in the correct place. Dangerous cargo must be stowed above deck, refrigerated containers (reefers) have their own section with power plugs, wine must be stowed far away from heat sources (engine room), etc.
- ensuring that the job of the operator at the next port isn't needlessly hard, i.e., not burying containers that will be unloaded soon.

All of this should be achieved while avoiding "wasted" container movements (any movement that isn't loading / unloading) as much as possible.

I don't have any experience at ports that are part of major trunk routes, but I would expect these issues to be slightly easier to manage for ships that essentially go back and forth between two major destinations (e.g., Shanghai - Long Beach). However, a port such as Long Beach also serves as a transport hub, so only one part of the containers unloaded are actually delivered at this port, with another substantial portion transferred to other ships.

Most customers pick up their empty containers a few days to a week before ETA, although in some cases, customers request earlier pick up and then use the containers as ad-hoc storage of their goods for a few days/weeks at a minor container yard, a train terminal, etc.

Once the stowage plan is made, the port can start receiving containers from customers. The containers are not simply stored in one big pile, but they must be kept separated by type / destination / weight / etc to ensure the stowage plan can be implemented efficiently. It is basically a big and very expensive version of [Tower of Hanoi](#). At this point, it starts getting problematic if a customer wants to make a change to their booking. Getting a container out from the middle of a stack will easily cause 50 container movements (billed at \$20 each), and the port doesn't necessarily have enough idle time with their machinery to even do this. A compromise here would

be to set this container aside when it naturally comes up in the loading sequence, which only costs one extra movement. This of course wastes time and makes operations harder.

The window for container arrivals closes around 48 hours before ETA, although sometimes the port will accept late arrivals for a fee. A certain percentage of containers always fails to show up, so a final round of adjustments has to be made to the stowage plan.

When the ship finally arrives, it will berth and immediately start unloading, and then loading. A typical stay of a ship is 1-2 days long. During this time, the cranes at the berth should be in continuous operation, not wasting a single second. Almost no amount of money will make the shipping company accept a change to a booking at this point.

After the ship leaves, we now have a second round of Towers of Hanoi at the port, this time with the inbound containers that were unloaded. These are stored at the port and handed out to importers as they send their trucks to pick them up.

Logistical Slack

The important thing to notice in port operations is that the container stacks are relatively close to being "random access", like the memory in your computer. Containers arrive and depart in an unpredictable manner, so they must be put into some sort of order to enable an efficient port stay for the ship. There are many easy suggestions for how to improve port logistics, but they often assume that containers are more or less fungible. This is a bad assumption.

If I was the benevolent dictator of Long Beach, I could enforce more efficient operations in a variety of ways. For example, I could just call in trucks, put arbitrary containers on those trucks, and tell the driver to go wherever this container must go. But not all truckers are willing to travel to arbitrary locations. Not all containers can accept the same kind of truck (e.g. due to weight differences). There are typically trade-offs in terms of reliability vs cost for different truckers. Also, each individual shipment (of one or a few containers) must be cleared by customs separately before it can leave the port.

Similarly, I could stack the containers much more closely together, sacrificing the "random access" property in the process, to create extra space in the port that might increase average throughput. However, this would destroy any guarantee that a given container will ever get out of the port. Humanity as a whole might be better off, but the owner of that particular cargo will be severely affected. The stacks must also be kept separate for different shipping companies. We are not at the point where we can make shipping itself fungible (e.g., forcing one company to carry the empty containers of another back to Asia) because this would require solving complex issues around the sharing of costs and liabilities.

What these kinds of proposals have in common is that they trade off a resource which I'll call "logistical slack" against an immediate, temporary relief of the congestion the port is facing. Losing this logistical slack is similar to "technical debt", a term from software engineering that describes how a piece of software becomes harder to modify and maintain over time, because the people working on it are incentivized to solve their short-term problems in fast and hacky ways, and neglect to maintain a

coherent architecture for the project. Technical debt is a bad place to be, and it tends to kill projects over time.

Giving up logistical slack is a terrible idea for very similar reasons. It would cause all operations to become more complex and expensive, like a debt that demands continuous interest payments. The momentary relief created would immediately be consumed, not necessarily in the most efficient way, but rather to appease those who complain the most forcefully. Furthermore, a port simply never has the downtime to do a "spring cleaning"-style reorganization. Thus, any problem created must be cleaned up incrementally over many, many cycles of normal operations. A decision that seems like a good idea now may leave a mess that lasts for years, with the true costs not becoming immediately apparent.

An Aside on Minor Container Yards

I'll now briefly focus on the container yards where the stacking limit of height 2 was famously removed in Long Beach. If I understand the situation correctly, these are minor yards mostly belonging to trucking companies, which are used as a base of operations for a fleet of trucks.

Such a yard will have some space that can serve as short-term storage for both full and empty containers. For example, a customer has a full warehouse, and wishes to send out a shipment before the shipping company opens its arrival window at the port. The trucking company can arrange for this using their own storage space. Another frequent situation is that there is a truck with an empty container that was just unloaded at some importer's facility. This container would usually be directly returned to the shipping company. However, if the truck is urgently needed, the trucker can save time by simply dropping it at their own yard, going on to do higher priority work, and return this container a few days later when they have some downtime and their opportunity cost is lower.

I believe that the stacking limit of two for these depots didn't arise by coincidence. It just happens to be that containers can be stacked up to two high with a standard forklift. This is exactly as sketchy as it sounds and has a tendency to damage the container and can even cause accidents. Any stacking more sophisticated than this requires specialized vehicles (reach loaders, top loaders, etc) that cannot easily travel and are thus dedicated to one site.

I don't want to claim expertise that I don't have, and I certainly don't know the specific situation in Long Beach. But for whatever it's worth, I believe quite strongly that no significant storage capacity was unlocked by Long Beach's suspension of stacking height limits.

Container Fleet Management

Container traffic is not balanced across both directions of each given trade route. Most noticeably, the flow of consumer goods from Asia to the US and Europe creates a huge imbalance, which the shipping companies correct for by making regular shipments of empty containers. Similarly, many locations have seasonal products requiring specialized containers. For example in Chile, there are yearly spikes in the demand for reefers, to export avocados and many other fruits and vegetables.

An important part of my job was to keep on top of these imbalances and plan for future demand, requesting empty shipments in and out of "my" ports, keeping the stock low, but not too low. Sometimes, reality did not match the sales projections, and then we ended up either with severe shortages or huge mountains of idle containers.

The former situation caused salespeople to be upset with me, and forced me to burn a lot of accumulated trust with my suppliers to make things happen on tighter timelines. Containers arriving on a given ship can in theory be turned around fast enough to be exported on the next ship just a week later, but this will require calls to importers pleading to return the containers fast, night shifts at the container depot for repairs, longer arrival windows at the port, etc. It can be done, but it's not a sustainable way to work. Any shipping company that tries to cut corners in this way will quickly find itself not receiving the favors (both small and large) that are needed to ensure smooth operations.

The latter situation (too much stock) caused the global offices to be upset with me, since containers are always in short supply somewhere in the world. There's nothing worse than asking for empty containers to be shipped to us, then sending them back empty four months later, with no paying shipment to cover the costs. In those cases we'd ask sales to call up their customers and offer below market rates to destinations X,Y,Z as a one-time deal, for big volumes only.

Shipping companies often rent large batches of containers on a long-term basis, with somewhat flexible conditions of pick up and return. This is a further tool used to balance the fleet. If bookings are lower than expected, the global office will simply return a few hundred containers in Asia to their owners instead of shipping them to us in Chile, while we use our excess stock to cover those future shipments. That's another source of non-fungibility, since different containers should preferentially be used on different routes, to ensure proximity to their eventual place of return. Again, there's no worse failure than having to ship an empty container halfway across the world with multiple transfers, only to return it to its owner.

I'll close this post with a funny anecdote. One of our depots once lost a container. Yes, they just *lost* an object measuring ~80 cubic meters. When it arrived, they entered it into their system, but the container was later not to be found in the stacks where it was supposed to be. They went and physically searched through the whole depot. There was a months long back and forth where we settled on them just having to buy the container from us, and doing whatever they wanted with it if it ever turned up. But then, just before finalizing this, I got a call from my contact at the depot. He was staring out his office window when he noticed that one of the containers that was visible in a stack just a few dozen meters away had a serial number that seemed strangely familiar... he looked up that number only to realize that it was in fact the long-lost container!

Ngo and Yudkowsky on alignment difficulty

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

This post is the first in a series of transcribed Discord conversations between Richard Ngo and Eliezer Yudkowsky, moderated by Nate Soares. We've also added Richard and Nate's running summaries of the conversation (and others' replies) from Google Docs.

Later conversation participants include Ajeya Cotra, Beth Barnes, Carl Shulman, Holden Karnofsky, Jaan Tallinn, Paul Christiano, Rob Bensinger, and Rohin Shah.

The transcripts are a complete record of several Discord channels MIRI made for discussion. We tried to edit the transcripts as little as possible, other than to fix typos and a handful of confusingly-worded sentences, to add some paragraph breaks, and to add referenced figures and links. We didn't end up redacting any substantive content, other than the names of people who would prefer not to be cited. We swapped the order of some chat messages for clarity and conversational flow (indicated with extra timestamps), and in some cases combined logs where the conversation switched channels.

Color key:

Chat by Richard and Eliezer	Other chat	Google Doc content	Inline comments
-----------------------------	------------	--------------------	-----------------

0. Prefatory comments

[Yudkowsky][8:32] (Nov. 6 follow-up comment)

(At Rob's request I'll try to keep this brief, but this was an experimental format and some issues cropped up that seem large enough to deserve notes.)

Especially when coming in to the early parts of this dialogue, I had some backed-up hypotheses about "What might be the main sticking point? and how can I address that?" which from the standpoint of a pure dialogue might seem to be causing me to go on digressions, relative to if I was just trying to answer Richard's own questions. On reading the dialogue, I notice that this looks evasive or like point-missing, like I'm weirdly not just directly answering Richard's questions.

Often the questions are answered later, or at least I think they are, though it may not be in the first segment of the dialogue. But the larger phenomenon is that I came in with some things I wanted to say, and Richard came in asking questions, and there was a minor accidental mismatch there. It would have looked better if we'd both stated positions first without question marks, say, or if I'd just confined myself to answering questions from Richard. (This is not a huge catastrophe, but it's something for the reader to keep in mind as a minor hiccup that showed up in the early parts of experimenting with this new format.)

[Yudkowsky][8:32] (Nov. 6 follow-up comment)

(Prompted by some later stumbles in attempts to summarize this dialogue. Summaries seem plausibly a major mode of propagation for a sprawling dialogue like this, and the following request seems like it needs to be very prominent to work - embedded requests later on didn't work.)

Please don't summarize this dialogue by saying, "and so Eliezer's MAIN idea is that" or "and then Eliezer thinks THE KEY POINT is that" or "the PRIMARY argument is that" etcetera. From my perspective, everybody comes in with a different set of sticking points versus things they see as obvious, and the conversation I have changes drastically depending on that. In the old days this used to be the Orthogonality Thesis, Instrumental Convergence, and superintelligence being a possible thing at all; today most OpenPhil-adjacent folks have other sticking points instead.

Please transform:

- "Eliezer's main reply is..." -> "Eliezer replied that..."
- "Eliezer thinks the key point is..." -> "Eliezer's point in response was..."
- "Eliezer thinks a major issue is..." -> "Eliezer replied that one issue is..."
- "Eliezer's primary argument against this is..." -> "Eliezer tried the counterargument that..."
- "Eliezer's main scenario for this is..." -> "In a conversation in September of 2021, Eliezer sketched a hypothetical where..."

Note also that the transformed statements say what you *observed*, whereas the untransformed statements are (often incorrect) *inferences* about my latent state of mind.

(Though "distinguishing relatively unreliable inference from more reliable observation" is not necessarily *the key idea here or the one big reason* I'm asking for this. That's just one point I tried making - one argument that I hope might help drive home the larger thesis.)

1. September 5 conversation

1.1. Deep vs. shallow problem-solving patterns

[Ngo][11:00]

Hi all! Looking forward to the discussion.

[Yudkowsky][11:01]

Hi and welcome all. My name is Eliezer and I think alignment is really actually quite extremely difficult. Some people seem to not think this! It's an important issue so ought to be resolved somehow, which we can hopefully fully do today. (I will however want to take a break after the first 90 minutes, if it goes that far and if Ngo is in sleep-cycle shape to continue past that.)

[Ngo][11:02]

A break in 90 minutes or so sounds good.

Here's one way to kick things off: I agree that humans trying to align arbitrarily capable AIs seems very difficult. One reason that I'm more optimistic (or at least, not confident that we'll have to face the full very difficult version of the problem) is that at a certain point AIs will be doing most of the work.

When you talk about alignment being difficult, what types of AIs are you thinking about aligning?

[Yudkowsky][11:04]

On my model of the Other Person, a lot of times when somebody thinks alignment shouldn't be that hard, they think there's some particular thing you can do to align an AGI, which isn't that hard, and their model is missing one of the foundational difficulties for why you can't do (easily or at all) one step of their procedure. So one of my own conversational processes might be to poke around looking for a step that the other person doesn't realize is hard. That said, I'll try to directly answer your own question first.

[Ngo][11:07]

I don't think I'm confident that there's any particular thing you can do to align an AGI. Instead I feel fairly uncertain over a broad range of possibilities for how hard the problem turns out to be.

And on some of the most important variables, it seems like evidence from the last decade pushes towards updating that the problem will be easier.

[Yudkowsky][11:09]

I think that after AGI becomes possible at all and then possible to scale to dangerously superhuman levels, there will be, in the best-case scenario where a lot of other social difficulties got resolved, a 3-month to 2-year period where only a very few actors have AGI, meaning that it was socially possible for those few actors to decide to *not* just scale it to where it automatically destroys the world.

During this step, if humanity is to survive, somebody has to perform some feat that causes the world to *not* be destroyed in 3 months or 2 years when too many actors have access to AGI code that will destroy the world if its intelligence dial is turned up. This requires that the first actor or actors to build AGI, be able to do *something* with that AGI which prevents the world from being destroyed; if it didn't require superintelligence, we could go do that thing right now, but no such human-doable act apparently exists so far as I can tell.

So we want the least dangerous, most easily aligned thing-to-do-with-an-AGI, but it does have to be a pretty powerful act to prevent the automatic destruction of Earth after 3 months or 2 years. It has to "flip the gameboard" rather than letting the suicidal game play out. We need to align the AGI that performs this pivotal act, to perform that pivotal act without killing everybody.

Parenthetically, no act powerful enough and gameboard-flipping enough to qualify is inside the Overton Window of politics, or possibly even of effective altruism, which presents a separate social problem. I usually dodge around this problem by picking an exemplar act which is powerful enough to actually flip the gameboard, but not the most alignable act because it would require way too many aligned details: Build self-replicating open-air nanosystems and use them (only) to melt all GPUs.

Since any such nanosystems would have to operate in the full open world containing lots of complicated details, this would require tons and tons of alignment work, is not the pivotal act easiest to align, and we should do some other thing instead. But the other thing I have in mind is also outside the Overton Window, just like this is. So I use "melt all GPUs" to talk about the requisite power level and the Overton Window problem level, both of which seem around the right levels to me, but the actual thing I have in mind is more alignable; and this way, I can reply to anyone who says "How dare you?!" by saying "Don't worry, I don't actually plan on doing that."

[Ngo][11:14]

One way that we could take this discussion is by discussing the pivotal act "make progress on the alignment problem faster than humans can".

[Yudkowsky][11:15]

This sounds to me like it requires extreme levels of alignment and operating in extremely dangerous regimes, such that, if you could do that, it would seem much more sensible to do some other pivotal act first, using a lower level of alignment tech.

[Ngo][11:16]

Okay, this seems like a crux on my end.

[Yudkowsky][11:16]

In particular, I would hope that - in unlikely cases where we survive at all - we were able to survive by operating a superintelligence only in the lethally dangerous, but still less dangerous, regime of "engineering nanosystems".

Whereas "solve alignment for us" seems to require operating in the even more dangerous regimes of "write AI code for us" and "model human psychology in tremendous detail".

[Ngo][11:17]

What makes these regimes so dangerous? Is it that it's very hard for humans to exercise oversight?

One thing that makes these regimes seem less dangerous to me is that they're broadly in the domain of "solving intellectual problems" rather than "achieving outcomes in the world".

[Yudkowsky][11:19][11:21]

Every AI output *effectuates* outcomes in the world. If you have a powerful unaligned mind hooked up to outputs that can start causal chains that effectuate dangerous things, it doesn't matter whether the comments on the code say "intellectual problems" or not.

The danger of "solving an intellectual problem" is when it requires a powerful mind to think about domains that, when solved, render very cognitively accessible strategies that can do dangerous things.

I expect the first alignment solution you can actually deploy in real life, in the unlikely event we get a solution at all, looks like 98% "don't think about all these topics that we do not absolutely need and are adjacent to

the capability to easily invent very dangerous outputs" and 2% "actually think about this dangerous topic but please don't come up with a strategy inside it that kills us".

[Ngo][11:21][11:22]

Let me try and be more precise about the distinction. It seems to me that systems which have been primarily trained to make predictions about the world would by default lack a lot of the cognitive machinery which humans use to take actions which pursue our goals.

Perhaps another way of phrasing my point is something like: it doesn't seem implausible to me that we build AIs that are significantly more intelligent (in the sense of being able to understand the world) than humans, but significantly less agentic.

Is this a crux for you?

(obviously "agentic" is quite underspecified here, so maybe it'd be useful to dig into that first)

[Yudkowsky][11:27][11:33]

I would certainly have learned very new and very exciting facts about intelligence, facts which indeed contradict my present model of how intelligences liable to be discovered by present research paradigms work, if you showed me... how can I put this in a properly general way... that problems I thought were about searching for states that get fed into a result function and then a result-scoring function, such that the input gets an output with a high score, were in fact not about search problems like that. I have sometimes given more specific names to this problem setup, but I think people have become confused by the terms I usually use, which is why I'm dancing around them.

In particular, just as I have a model of the Other Person's Beliefs in which they think alignment is easy because they don't know about difficulties I see as very deep and fundamental and hard to avoid, I also have a model in which people think "why not just build an AI which does X but not Y?" because they don't realize what X and Y have in common, which is something that draws deeply on having deep models of intelligence. And it is hard to convey this deep theoretical grasp.

But you can also see powerful practical hints that these things are much more correlated than, eg, Robin Hanson was imagining during the [FOOM debate](#), because Robin did not think something like GPT-3 should exist; Robin thought you should need to train lots of specific domains that didn't generalize. I argued then with Robin that it was something of a hint that humans had visual cortex and cerebellar cortex but not Car Design Cortex, in order to design cars. Then in real life, it proved that reality was far to the Eliezer side of Eliezer on the [Eliezer-Robin axis](#), and things like GPT-3 were built with *less* architectural complexity and generalized *more*

than I was arguing to Robin that complex architectures should generalize over domains.

The metaphor I sometimes use is that it is very hard to build a system that drives cars painted red, but is not at all adjacent to a system that could, with a few alterations, prove to be very good at driving a car painted blue. The "drive a red car" problem and the "drive a blue car" problem have too much in common. You can maybe ask, "Align a system so that it has the capability to drive red cars, but refuses to drive blue cars." You can't make a system that is very good at driving red-painted cars, but lacks the basic capability to drive blue-painted cars because you never trained it on that. The patterns found by gradient descent, by genetic algorithms, or by other plausible methods of optimization, for driving red cars, would be patterns very close to the ones needed to drive blue cars. When you optimize for red cars you get the blue car *capability* whether you like it or not.

[Ngo][11:32]

Does your model of intelligence rule out building AIs which make dramatic progress in mathematics without killing us all?

[Yudkowsky][11:34][11:39]

If it were possible to perform some pivotal act that saved the world with an AI that just made progress on proving mathematical theorems, without, eg, needing to explain those theorems to humans, I'd be *extremely* interested in that as a potential pivotal act. We wouldn't be out of the woods, and I wouldn't actually know how to build an AI like that without killing everybody, but it would immediately trump everything else as the obvious line of research to pursue.

Parenthetically, there is very very little which my model of intelligence *rules out*. I think we all die because we cannot do certain dangerous things correctly, *on the very first try in the dangerous regimes where one mistake kills you*, and do them *before* proliferation of much easier technologies kills us. If you have the Textbook From 100 Years In The Future that gives the simple robust solutions for everything, that actually work, you can write a superintelligence that thinks $2 + 2 = 5$ because the Textbook gives the methods for doing that which are simple and actually work in practice in real life.

(The Textbook has the equivalent of "use ReLUs instead of sigmoids" everywhere, and avoids all the clever-sounding things that will work at subhuman levels and blow up when you run them at superintelligent levels.)

[Ngo][11:36][11:40]

Hmm, so suppose we train an AI to prove mathematical theorems when given them, perhaps via some sort of adversarial setter-solver training process.

By default I have the intuition that this AI could become extremely good at proving theorems - far beyond human level - without having goals about real-world outcomes.

It seems to me that in your model of intelligence, being able to do tasks like mathematics is closely coupled with trying to achieve real-world outcomes. But I'd actually take GPT-3 as some evidence against this position (although still evidence in favour of your position over Hanson's), since it seems able to do a bunch of reasoning tasks while still not being very agentic.

There's some alternative world where we weren't able to train language models to do reasoning tasks without first training them to perform tasks in complex RL environments, and in that world I'd be significantly less optimistic.

[Yudkowsky][11:41]

I put to you that there is a predictable bias in your estimates, where you don't know about the Deep Stuff that is required to prove theorems, so you imagine that certain cognitive capabilities are more disjoint than they actually are. If you knew about the things that humans are using to reuse their reasoning about chipped handaxes and other humans, to prove math theorems, you would see it as more plausible that proving math theorems would generalize to chipping handaxes and manipulating humans.

GPT-3 is a... complicated story, on my view of it and intelligence. We're looking at an interaction between tons and tons of memorized shallow patterns. GPT-3 is very unlike the way that natural selection built humans.

[Ngo][11:44]

I agree with that last point. But this is also one of the reasons that I previously claimed that AIs could be more intelligent than humans while being less agentic, because there are systematic differences between the way in which natural selection built humans, and the way in which we'll train AGIs.

[Yudkowsky][11:45]

My current suspicion is that Stack More Layers alone is not going to take us to GPT-6 which is a true AGI; and this is because of the way that GPT-3 is, in your own terminology, "not agentic", and which is, in my terminology, not having gradient descent on GPT-3 run across sufficiently deep problem-solving patterns.

[Ngo][11:46]

Okay, that helps me understand your position better.

So here's one important difference between humans and neural networks: humans face the genomic bottleneck which means that each individual has to rederive all the knowledge about the world that their parents already had. If this genetic bottleneck hadn't been so tight, then individual humans would have been significantly less capable of performing novel tasks.

[Yudkowsky][11:50]

I agree.

[Ngo][11:50]

In my terminology, this is a reason that humans are "more agentic" than we otherwise would have been.

[Yudkowsky][11:50]

This seems indisputable.

[Ngo][11:51]

Another important difference: humans were trained in environments where we had to run around surviving all day, rather than solving maths problems etc.

[Yudkowsky][11:51]

I continue to nod.

[Ngo][11:52]

Supposing I agree that reaching a certain level of intelligence will require AIs with the "deep problem-solving patterns" you talk about, which lead AIs to try to achieve real-world goals. It still seems to me that there's likely a lot of space between that level of intelligence, and human intelligence.

And if that's the case, then we could build AIs which help us solve the alignment problem before we build AIs which instantiate sufficiently deep problem-solving patterns that they decide to take over the world.

Nor does it seem like the reason *humans* want to take over the world is because of a deep fact about our intelligence. It seems to me that

humans want to take over the world mainly because that's very similar to things we evolved to do (like taking over our tribe).

[Yudkowsky][11:57]

So here's the part that I agree with: If there were one theorem only mildly far out of human reach, like proving the ABC Conjecture (if you think it hasn't already been proven), and providing a machine-readable proof of this theorem would immediately save the world - say, aliens will give us an aligned superintelligence, as soon as we provide them with this machine-readable proof - then there would exist a plausible though not certain road to saving the world, which would be to try to build a *shallow* mind that proved the ABC Conjecture by memorizing tons of relatively shallow patterns for mathematical proofs learned through self-play; without that system ever abstracting math as deeply as humans do, but the sheer width of memory and sheer depth of search sufficing to do the job. I am not sure, to be clear, that this would work. But my model of intelligence does not rule it out.

[Ngo][11:58]

(I'm actually thinking of a mind which understands maths more deeply than humans - but perhaps only understands maths, or perhaps also a range of other sciences better than humans.)

[Yudkowsky][12:00]

Parts I disagree with: That "help us solve alignment" bears any significant overlap with "provide us a machine-readable proof of the ABC Conjecture without thinking too deeply about it". That humans want to take over the world only because it resembles things we evolved to do.

[Ngo][12:01]

I definitely agree that humans don't *only* want to take over the world because it resembles things we evolved to do.

[Yudkowsky][12:02]

Alas, eliminating 5 reasons why something would go wrong doesn't help much if there's 2 remaining reasons something would go wrong that are much harder to eliminate!

[Ngo][12:02]

But if we imagine having a human-level intelligence which *hadn't* evolved primarily to do things that reasonably closely resembled taking over the

world, then I expect that we could ask that intelligence questions in a fairly safe way.

And that's also true for an intelligence that is noticeably above human level.

So one question is: how far above human level could we get before a system which has only been trained to do things like answer questions and understand the world will decide to take over the world?

[Yudkowsky][12:04]

I think this is one of the very rare cases where the intelligence difference between "village idiot" and "Einstein", which I'd usually see as very narrow, makes a structural difference! I think you can get some outputs from a village-idiot-level AGI, which got there by training on domains exclusively like math, and this will probably not destroy the world (*if* you were right about that, about what was going on inside). I have more concern about the Einstein level.

[Ngo][12:05]

Let's focus on the Einstein level then.

Human brains have been optimised very little for doing science.

This suggests that building an AI which is Einstein-level at doing science is significantly easier than building an AI which is Einstein-level at taking over the world (or other things which humans evolved to do).

[Yudkowsky][12:08]

I think there's a certain broad sense in which I agree with the literal truth of what you just said. You will systematically overestimate *how much* easier, or how far you can push the science part without getting the taking-over-the-world part, for as long as your model is ignorant of what they have in common.

[Ngo][12:08]

Maybe this is a good time to dig into the details of what they have in common, then.

[Yudkowsky][12:09][12:11]][12:13]

I feel like I haven't had much luck with trying to explain that on previous occasions. Not to you, to others too.

There are shallow topics like why p-zombies can't be real and how quantum mechanics works and why science ought to be using likelihood functions instead of p-values, and I can *barely* explain those to *some* people, but then there are some things that are apparently much harder to explain than that and which defeat my abilities as an explainer.

That's why I've been trying to point out that, even if you don't know the specifics, there's an estimation bias that you can realize should exist in principle.

Of course, I also haven't had much luck in saying to people, "Well, even if you don't know the truth about X that would let you see Y, can you not see by abstract reasoning that knowing *any* truth about X would predictably cause you to update in the direction of Y" - people don't seem to actually internalize that much either. Not you, other discussions.

[Ngo][12:10][12:11][12:13]

Makes sense. Are there ways that I could try to make this easier? E.g. I could do my best to explain what I think your position is.

Given what you've said I'm not optimistic about this helping much.

But insofar as this is the key set of intuitions which has been informing your responses, it seems worth a shot.

Another approach would be to focus on our predictions for how AI capabilities will play out over the next few years.

I take your point about my estimation bias. To me it feels like there's also a bias going the other way, which is that as long as we don't know the mechanisms by which different human capabilities work, we'll tend to lump them together as one thing.

[Yudkowsky][12:14]

Yup. If you didn't know about visual cortex and auditory cortex, or about eyes and ears, you would assume much more that any sentience ought to both see and hear.

[Ngo][12:16]

So then my position is something like: human pursuit of goals is driven by emotions and reward signals which are deeply evolutionarily ingrained, and without those we'd be much safer but not that much worse at pattern recognition.

[Yudkowsky][12:17]

If there's a pivotal act you can get just by supreme acts of pattern recognition, that's right up there with "pivotal act composed solely of math" for things that would obviously instantly become the prime direction of research.

[Ngo][12:18]

To me it seems like maths is *much more* about pattern recognition than, say, being a CEO. Being a CEO requires coherence over long periods of time; long-term memory; motivation; metacognition; etc.

[Yudkowsky][12:18][12:23]

(One occasionally-argued line of research can be summarized from a certain standpoint as "how about a pivotal act composed entirely of predicting text" and to this my reply is "you're trying to get fully general AGI capabilities by predicting text that is *about* deep / 'agentic' reasoning, and that doesn't actually help".)

Human math is very much about goals. People want to prove subtheorems on the way to proving theorems. We might be able to make a *different* kind of mathematician that works more like GPT-3 in the dangerously inscrutable parts that are all noninspectable vectors of floating-point numbers, but even there you'd need some Alpha-Zero-like outer framework to supply the direction of search.

That outer framework might be able to be powerful enough without being reflective, though. So it would plausibly be *much easier* to build a mathematician that was capable of superhuman formal theorem-proving but not agentic. The reality of the world might tell us "lolnope" but my model of intelligence doesn't mandate that. That's why, if you gave me a pivotal act composed entirely of "output a machine-readable proof of this theorem and the world is saved", I would pivot there! It actually does seem like it would be a lot easier!

[Ngo][12:21][12:25]

Okay, so if I attempt to rephrase your argument:

Your position: There's a set of fundamental similarities between tasks like doing maths, doing alignment research, and taking over the world. In all of these cases, agents based on techniques similar to modern ML which are very good at them will need to make use of deep problem-solving patterns which include goal-oriented reasoning. So while it's possible to beat humans at some of these tasks without those core competencies, people usually overestimate the extent to which that's possible.

[Yudkowsky][12:25]

Remember, a lot of my concern is about what happens *first*, especially if it happens soon enough that future AGI bears any resemblance whatsoever to modern ML; not about what can be done in principle.

[Soares][12:26]

(Note: it's been 85 min, and we're planning to take a break at 90min, so this seems like a good point for a little bit more clarifying back-and-forth on Richard's summary before a break.)

[Ngo][12:26]

I'll edit to say "plausible for ML techniques"?

(and "extent to which that's plausible")

[Yudkowsky][12:28]

I think that obvious-to-me future outgrowths of modern ML paradigms are *extremely* liable to, if they can learn how to do sufficiently superhuman X, generalize to taking over the world. How fast this happens does depend on X. It would plausibly happen relatively slower (at higher levels) with theorem-proving as the X, and with architectures that carefully stuck to gradient-descent-memorization over shallow network architectures to do a pattern-recognition part with search factored out (sort of, this is not generally safe, this is not a general formula for safe things!); rather than imposing anything like the genetic bottleneck you validly pointed out as a reason why humans generalize. Profitable X, and all X I can think of that would actually save the world, seem much more problematic.

[Ngo][12:30]

Okay, happy to take a break here.

[Soares][12:30]

Great timing!

[Ngo][12:30]

We can do a bit of meta discussion afterwards; my initial instinct is to push on the question of how similar Eliezer thinks alignment research is to theorem-proving.

[Yudkowsky][12:30]

Yup. This is my lunch break (actually my first-food-of-day break on a 600-calorie diet) so I can be back in 45min if you're still up for that.

[Ngo][12:31]

Sure.

Also, if any of the spectators are reading in real time, and have suggestions or comments, I'd be interested in hearing them.

[Yudkowsky][12:31]

I'm also cheerful about spectators posting suggestions or comments during the break.

[Soares][12:32]

Sounds good. I declare us on a break for 45min, at which point we'll reconvene (for another 90, by default).

Floor's open to suggestions & commentary.

1.2. Requirements for science

[Yudkowsky][12:50]

I seem to be done early if people (mainly Richard) want to resume in 10min (30m break)

[Ngo][12:51]

Yepp, happy to do so

[Soares][12:57]

Some quick commentary from me:

- It seems to me like we're exploring a crux in the vicinity of "should we expect that systems capable of executing a pivotal act would, by default in lieu of significant technical alignment effort, be using their outputs to optimize the future".

- I'm curious whether you two agree that this is a crux (but plz don't get side-tracked answering me).
- The general discussion seems to be going well to me.
 - In particular, huzzah for careful and articulate efforts to zero in on cruxes.

[Ngo][13:00]

I think that's a crux for the specific pivotal act of "doing better alignment research", and maybe some other pivotal acts, but not all (or necessarily most) of them.

[Yudkowsky][13:01]

I should also say out loud that I've been working a bit with Ajeya on making an attempt to convey the intuitions behind there being deep patterns that generalize and are liable to be learned, which covered a bunch of ground, taught me how much ground there was, and made me relatively more reluctant to try to re-cover the same ground in this modality.

[Ngo][13:02]

Going forward, a couple of things I'd like to ask Eliezer about:

- In what ways are the tasks that are most useful for alignment similar or different to proving mathematical theorems (which we agreed might generalise relatively slowly to taking over the world)?
- What are the deep problem-solving patterns underlying these tasks?
- Can you summarise my position?

I was going to say that I was most optimistic about #2 in order to get these ideas into a public format

But if that's going to happen anyway based on Ajeya's work, then that seems less important

[Yudkowsky][13:03]

I could still try briefly and see what happens.

[Ngo][13:03]

That seems valuable to me, if you're up for it.

At the same time, I'll try to summarise some of my own intuitions about intelligence which I expect to be relevant.

[Yudkowsky][13:04]

I'm not sure I could summarize your position in a non-straw way. To me there's a huge visible distance between "solve alignment for us" and "output machine-readable proofs of theorems" where I can't give a good account of why you think talking about the latter would tell us much about the former. I don't know what other pivotal act you think might be easier.

[Ngo][13:06]

I see. I was considering "solving scientific problems" as an alternative to "proving theorems", with alignment being one (particularly hard) example of a scientific problem.

But decided to start by discussing theorem-proving since it seemed like a clearer-cut case.

[Yudkowsky][13:07]

Can you predict in advance why Eliezer thinks "solving scientific problems" is significantly thornier? (Where alignment is like totally not "a particularly hard example of a scientific problem" except in the sense that it has science in it at all; which is maybe the real crux; but also a more difficult issue.)

[Ngo][13:09]

Based on some of your earlier comments, I'm currently predicting that you think the step where the solutions need to be legible to and judged by humans makes science much thornier than theorem-proving, where the solutions are machine-checkable.

[Yudkowsky][13:10]

That's one factor. Should I state the other big one or would you rather try to state it first?

[Ngo][13:10]

Requiring a lot of real-world knowledge for science?

If it's not that, go ahead and say it.

[Yudkowsky][13:11]

That's one way of stating it. The way I'd put it is that it's about making up hypotheses about the real world.

Like, the real world is then a thing that the AI is modeling, at all.

Factor 3: On many interpretations of doing science, you would furthermore need to think up experiments. That's planning, value-of-information, search for an experimental setup whose consequences distinguish between hypotheses (meaning you're now searching for initial setups that have particular causal consequences).

[Ngo][13:12]

To me "modelling the real world" is a very continuous variable. At one end you have physics equations that are barely separable from maths problems, at the other end you have humans running around in physical bodies.

To me it seems plausible that we could build an agent which solves scientific problems but has very little self-awareness (in the sense of knowing that it's an AI, knowing that it's being trained, etc).

I expect that your response to this is that modelling oneself is part of the deep problem-solving patterns which AGIs are very likely to have.

[Yudkowsky][13:15]

There's a problem of *inferring the causes of sensory experience* in cognition-that-does-science. (Which, in fact, also appears in the way that humans do math, and is possibly inextricable from math in general; but this is an example of the sort of deep model that says "Whoops I guess you get science from math after all", not a thing that makes science less dangerous because it's more like just math.)

You can build an AI that only ever drives red cars, and which, at no point in the process of driving a red car, ever needs to drive a blue car in order to drive a red car. That doesn't mean its red-car-driving capabilities won't be extremely close to blue-car-driving capabilities if at any point the internal cognition happens to get pointed towards driving a blue car.

The fact that there's a deep car-driving pattern which is the same across red cars and blue cars doesn't mean that the AI has ever driven a blue car, per se, or that it has to drive blue cars to drive red cars. But if blue cars are fire, you sure are playing with that fire.

[Ngo][13:18]

To me, "sensory experience" as in "the video and audio coming in from this body that I'm piloting" and "sensory experience" as in "a file containing the most recent results of the large hadron collider" are very very different.

(I'm not saying we could train an AI scientist just from the latter - but plausibly from data that's closer to the latter than the former)

[Yudkowsky][13:19]

So there's separate questions about "does an AGI *inseparably need* to model itself inside the world to do science" and "did we build something that would be very close to modeling itself, and could easily stumble across that by accident somewhere in the inscrutable floating-point numbers, especially if that was even slightly useful for solving the outer problems".

[Ngo][13:19]

Hmm, I see

[Yudkowsky][13:20][13:21][13:21]

If you're trying to build an AI that literally does science only to observations collected without the AI having had a causal impact on those observations, that's legitimately "more dangerous than math but maybe less dangerous than active science".

You might still stumble across an active scientist because it was a simple internal solution to something, but the outer problem would be legitimately stripped of an important structural property the same way that pure math not describing Earthly objects is stripped of important structural properties.

And of course my reaction again is, "There is no pivotal act which uses only that cognitive capability."

[Ngo][13:20][13:21][13:26]

I guess that my (fairly strong) prior here is that something like self-modelling, which is very deeply built into basically every organism, is a very hard thing for an AI to stumble across by accident without significant optimisation pressure in that direction.

But I'm not sure how to argue this except by digging into your views on what the deep problem-solving patterns are. So if you're still willing to briefly try and explain those, that'd be useful to me.

"Causal impact" again seems like a very continuous variable - it seems like the *amount* of causal impact you need to do good science is much less than the amount which is needed to, say, be a CEO.

[Yudkowsky][13:26]

The amount doesn't seem like the key thing, nearly so much as what underlying facilities you need to do whatever amount of it you need.

[Ngo][13:27]

Agreed.

[Yudkowsky][13:27]

If you go back to the 16th century and ask for just one mRNA vaccine, that's not much of a difference from asking for a million hundred of them.

[Ngo][13:28]

Right, so the additional premise which I'm using here is that the ability to reason about causally impacting the world in order to achieve goals is something that you can have a little bit of.

Or a lot of, and that the difference between these might come down to the training data used.

Which at this point I don't expect you to agree with.

[Yudkowsky][13:29]

If you have reduced a pivotal act to "look over the data from this hadron collider you neither built nor ran yourself", that really is a structural step down from "do science" or "build a nanomachine". But I can't see any pivotal acts like that, so is that question much of a crux?

If there's intermediate steps they might be described in my native language like "reason about causal impacts across only this one preprogrammed domain which you didn't learn in a general way, in only this part of the cognitive architecture that is separable from the rest of the cognitive architecture".

[Ngo][13:31]

Perhaps another way of phrasing this intermediate step is that the agent has a shallow understanding of how to induce causal impacts.

[Yudkowsky][13:31]

What is "shallow" to you?

[Ngo][13:31]

In a similar way to how you claim that GPT-3 has a shallow understanding of language.

[Yudkowsky][13:32]

So it's memorized a ton of shallow causal-impact-inducing patterns from a large dataset, and this can be verified by, for example, presenting it with an example mildly outside the dataset and watching it fail, which we think will confirm our hypothesis that it didn't learn any deep ways of solving that dataset.

[Ngo][13:33]

Roughly speaking, yes.

[Yudkowsky][13:34]

Eg, it wouldn't surprise us at all if GPT-4 had learned to predict "27 * 18" but not "what is the area of a rectangle 27 meters by 18 meters"... is what I'd like to say, but Codex sure did demonstrate those two were kinda awfully proximal.

[Ngo][13:34]

Here's one way we could flesh this out. Imagine an agent that loses coherence quickly when it's trying to act in the world.

So for example, we've trained it to do scientific experiments over a period of a few hours or days

And then it's very good at understanding the experimental data and extracting patterns from it

But upon running it for a week or a month, it loses coherence in a similar way to how GPT-3 loses coherence - e.g. it forgets what it's doing.

My story for why this might happen is something like: there is a specific skill of having long-term memory, and we never trained our agent to have this skill, and so it has not acquired that skill (even though it can reason in very general and powerful ways in the short term).

This feels similar to the argument I was making before about how an agent might lack self-awareness, if we haven't trained it specifically to have that.

[Yudkowsky][13:39]

There's a set of obvious-to-me tactics for doing a pivotal act with minimal danger, which I do not think collectively make the problem safe, and one

of these sets of tactics is indeed "Put a limit on the 'attention window' or some other internal parameter, ramp it up slowly, don't ramp it any higher than you needed to solve the problem."

[Ngo][13:41]

You could indeed do this manually, but my expectation is that you could also do this automatically, by training agents in environments where they don't benefit from having long attention spans.

[Yudkowsky][13:42]

(Any time one imagines a specific tactic of this kind, if one has the [security mindset](#), one can also imagine all sorts of ways it might go wrong; for example, an attention window can be defeated if there's any aspect of the attended data or the internal state that ended up depending on past events in a way that leaked info about them. But, depending on how much superintelligence you were throwing around elsewhere, you could maybe get away with that, some of the time.)

[Ngo][13:43]

And that if you put agents in environments where they answer questions but don't interact much with the physical world, then there will be many different traits which are necessary for achieving goals in the real world which they will lack, because there was little advantage to the optimiser of building those traits in.

[Yudkowsky][13:43]

I'll observe that TransformerXL built an attention window that generalized, trained it on I think 380 tokens or something like that, and then found that it generalized to 4000 tokens or something like that.

[Ngo][13:43]

Yeah, an order of magnitude of generalisation is not surprising to me.

[Yudkowsky][13:44]

Having observed one order of magnitude, I would personally not be surprised by two orders of magnitude either, after seeing that.

[Ngo][13:45]

I'd be a little surprised, but I assume it would happen eventually.

1.3. Capability dials

[Yudkowsky][13:46]

I have a sense that this is all circling back to the question, "But what is it we *do* with the intelligence thus weakened?" If you can save the world using a rock, I can build you a very safe rock.

[Ngo][13:46]

Right.

So far I've said "alignment research", but I haven't been very specific about it.

I guess some context here is that I expect that the first things we do with intelligence similar to this is create great wealth, produce a bunch of useful scientific advances, etc.

And that we'll be in a world where people take the prospect of AGI much more seriously

[Yudkowsky][13:48]

I mostly expect - albeit with some chance that reality says "So what?" to me and surprises me, because it is not as solidly determined as some other things - that we do not hang around very long in the "weirdly ~human AGI" phase before we get into the "if you crank up this AGI it destroys the world" phase. Less than 5 years, say, to put numbers on things.

It would not surprise me in the least if the world ends before self-driving cars are sold on the mass market. On some quite plausible scenarios which I think have >50% of my probability mass at the moment, research AGI companies would be able to produce prototype car-driving AIs if they spent time on that, given the near-world-ending tech level; but there will be Many Very Serious Questions about this relatively new unproven advancement in machine learning being turned loose on the roads. And their AGI tech will gain the property "can be turned up to destroy the world" before Earth gains the property "you're allowed to sell self-driving cars on the mass market" because there just won't be much time.

[Ngo][13:52]

Then I expect that another thing we do with this is produce a very large amount of data which rewards AIs for following human instructions.

[Yudkowsky][13:52]

On other scenarios, of course, self-driving becomes possible by limited AI well before things start to break (further) on AGI. And on some scenarios, the way you got to AGI was via some breakthrough that is already scaling pretty fast, so by the time you can use the tech to get self-driving cars, that tech already ends the world if you turn up the dial, or that event follows very swiftly.

[Ngo][13:53]

When you talk about "cranking up the AGI", what do you mean?

Using more compute on the same data?

[Yudkowsky][13:53]

Running it with larger bounds on the for loops, over more GPUs, to be concrete about it.

[Ngo][13:53]

In a RL setting, or a supervised, or unsupervised learning setting?

Also: can you elaborate on the for loops?

[Yudkowsky][13:56]

I do not quite think that gradient descent on Stack More Layers alone - as used by OpenAI for GPT-3, say, and as opposed to Deepmind which builds more complex artifacts like Mu Zero or AlphaFold 2 - is liable to be the first path taken to AGI. I am reluctant to speculate more in print about clever ways to AGI, and I think any clever person out there will, if they are really clever and not just a fancier kind of stupid, not talk either about what they think is missing from Stack More Layers or how you would really get AGI. That said, the way that you cannot just run GPT-3 at a greater search depth, the way you can run Mu Zero at a greater search depth, is part of why I think that AGI is not likely to look exactly like GPT-3; the thing that kills us is likely to be a thing that can get more dangerous when you turn up a dial on it, not a thing that intrinsically has no dials that can make it more dangerous.

1.4. Consequentialist goals vs. deontologist goals

[Ngo][13:59]

Hmm, okay. Let's take a quick step back and think about what would be useful for the last half hour.

I want to flag that my intuitions about pivotal acts are not very specific; I'm quite uncertain about how the geopolitics of that situation would work, as well as the timeframe between somewhere-near-human-level AGI and existential risk AGI.

So we could talk more about this, but I expect there'd be a lot of me saying "well we can't rule out that X happens", which is perhaps not the most productive mode of discourse.

A second option is digging into your intuitions about how cognition works.

[Yudkowsky][14:03]

Well, obviously, in the limit of alignment not being accessible to our civilization, and my successfully building a model weaker than reality which nonetheless correctly rules out alignment being accessible to our civilization, I could spend the rest of my short remaining lifetime arguing with people whose models are weak enough to induce some area of ignorance where for all they know you could align a thing. But that is predictably how conversations go in possible worlds where the Earth is doomed; so somebody wiser on the meta-level, though also ignorant on the object-level, might prefer to ask: "Where do you think your knowledge, rather than your ignorance, says that alignment ought to be doable and you will be surprised if it is not?"

[Ngo][14:07]

That's a fair point. Although it seems like a structural property of the "pivotal act" framing, which builds in doom by default.

[Yudkowsky][14:08]

We could talk about that, if you think it's a crux. Though I'm also not thinking that this whole conversation gets done in a day, so maybe for publishability reasons we should try to focus more on one line of discussion?

But I do think that lots of people get their optimism by supposing that the world can be saved by doing less dangerous things with an AGI. So it's a big ol' crux of mine on priors.

[Ngo][14:09]

Agreed that one line of discussion is better; I'm happy to work within the pivotal act framing for current purposes.

A third option is that I make some claims about how cognition works, and we see how much you agree with them.

[Yudkowsky][14:12]

(Though it's something of a restatement, a reason I'm not going into "my intuitions about how cognition works" is that past experience has led me to believe that conveying this info in a form that the Other Mind will actually absorb and operate, is really quite hard and takes a long discussion, relative to my current abilities to Actually Explain things; it is the sort of thing that might take doing homework exercises to grasp how one structure is appearing in many places, as opposed to just being flatly told that to no avail, and I have not figured out the homework exercises.)

I'm cheerful about hearing your own claims about cognition and disagreeing with them.

[Ngo][14:12]

Great

Okay, so one claim is that something like deontology is a fairly natural way for minds to operate.

[Yudkowsky][14:14]

("If that were true," he thought at once, "bureaucracies and books of regulations would be a lot more efficient than they are in real life.")

[Ngo][14:14]

Hmm, although I think this was probably not a very useful phrasing, let me think about how to rephrase it.

Okay, so in [our earlier email discussion](#), we talked about the concept of "obedience".

To me it seems like it is just as plausible for a mind to have a concept like "obedience" as its rough goal, as a concept like maximising paperclips.

If we imagine training an agent on a large amount of data which pointed in the rough direction of rewarding obedience, for example, then I imagine that by default obedience would be a constraint of comparable strength to, say, the human survival instinct.

(Which is obviously not strong enough to stop humans doing a bunch of things that contradict it - but it's a pretty good starting point.)

[Yudkowsky][14:18]

Heh. You mean of comparable strength to the human instinct to explicitly maximize inclusive genetic fitness?

[Ngo][14:19]

Genetic fitness wasn't a concept that our ancestors were able to understand, so it makes sense that they weren't pointed directly towards it.

(And nor did they understand *how* to achieve it.)

[Yudkowsky][14:19]

Even in that paradigm, except insofar as you expect gradient descent to work very differently from gene-search optimization - which, admittedly, it does - when you optimize really hard on a thing, you get contextual correlates to it, not the thing you optimized on.

This is of course one of the Big Fundamental Problems that I expect in alignment.

[Ngo][14:20]

Right, so the main correlate that I've seen discussed is "do what would make the human give you a high rating, not what the human actually wants"

One thing I'm curious about is the extent to which you're concerned about this specific correlate, versus correlates in general.

[Yudkowsky][14:21]

That said, I also see basic structural reasons why paperclips would be much easier to train than "obedience", even if we could magically instill simple inner desires that perfectly reflected the simple outer algorithm we saw ourselves as running over many particular instances of a loss function.

[Ngo][14:22]

I'd be interested in hearing what those are.

[Yudkowsky][14:22]

well, first of all, why *is* a book of regulations so much more unwieldy than a hunter-gatherer?

if deontology is just as good as [consequentialism](#), y'know.

(do you want to try replying or should I just say?)

[Ngo][14:23]

Go ahead

I should probably clarify that I agree that you can't just replace consequentialism with deontology

The claim is more like: when it comes to high-level concepts, it's not clear to me why high-level consequentialist goals are more natural than high-level deontological goals.

[Yudkowsky][14:24]

I reply that reality is complicated, so when you pump a simple goal through complicated reality you get complicated behaviors required to achieve the goal. If you think of reality as a complicated function Input->Probability(Output), then even to get a simple Output or a simple partition on Output or a high expected score in a simple function over Output, you may need very complicated Input.

Humans don't trust each other. They imagine, "Well, if I just give this bureaucrat a goal, perhaps they won't reason honestly about what it takes to achieve that goal! Oh no! Therefore I will instead, being the trustworthy and accurate person that I am, reason myself about constraints and requirements on the bureaucrat's actions, such that, if the bureaucrat obeys these regulations, I expect the outcome of their action will be what I want."

But (compared to a general intelligence that observes and models complicated reality and does its own search to pick actions) an actually-effective book of regulations (implemented by some nonhuman mind with a large enough and perfect enough memory to memorize it) would tend to involve a (physically unmanageable) vast number of rules saying "if you observe this, do that" to follow all the crinkles of complicated reality as it can be inferred from observation.

[Ngo][14:28]

(Though it's something of a restatement, a reason I'm not going into "my intuitions about how cognition works" is that past experience has led me to believe that conveying this info in a form that the Other Mind will actually absorb and operate, is really quite hard and takes a long discussion, relative to my current abilities to Actually Explain things; it is the sort of thing that might take doing homework exercises to grasp how one structure is appearing in many places, as opposed to just being flatly told that to no avail, and I have not figured out the homework exercises.)

(As a side note: do you have a rough guess for when your work with Ajeya will be made public? If it's still a while away, I'm wondering whether it's still useful to have a rough outline of these intuitions even if it's in a form that very few people will internalise)

[Yudkowsky][14:30]

(As a side note: do you have a rough guess for when your work with Ajeya will be made public? If it's still a while away, I'm wondering whether it's still useful to have a rough outline of these intuitions even if it's in a form that very few people will internalise)

Plausibly useful, but not to be attempted today, I think?

[Ngo][14:30]

Agreed.

[Yudkowsky][14:30]

(We are now theoretically in overtime, which is okay for me, but for you it is 11:30pm (I think?) and so it is on you to call when to halt, now or later.)

[Ngo][14:32]

Yeah, it's 11.30 for me. I think probably best to halt here. I agree with all the things you just said about reality being complicated, and why consequentialism is therefore valuable. My "deontology" claim (which was, in its original formulation, far too general - apologies for that) was originally intended as a way of poking into your intuitions about which types of cognition are natural or unnatural, which I think is the topic we've been circling around for a while.

[Yudkowsky][14:33]

Yup, and a place to resume next time might be why I think "obedience" is unnatural compared to "paperclips" - though that is a thing that probably requires taking that stab at what underlies surface competencies.

[Ngo][14:34]

Right. I do think that even a vague gesture at that would be reasonably helpful (assuming that this doesn't already exist online?)

[Yudkowsky][14:34]

Not yet afaik, and I don't want to point you to Ajeya's stuff even if she were ok with that, because then this in-context conversation won't make sense to others.

[Ngo][14:35]

For my part I should think more about pivotal acts that I'd be willing to specifically defend.

In any case, thanks for the discussion 😊

Let me know if there's a particular time that suits you for a follow-up; otherwise we can sort it out later.

[Soares][14:37]

(y'all are doing all my jobs for me)

[Yudkowsky][14:37]

could try Tuesday at this same time - though I may be in worse shape for dietary reasons, still, seems worth trying.

[Soares][14:37]

(wfm)

[Ngo][14:39]

Tuesday not ideal, any others work?

[Yudkowsky][14:39]

Wednesday?

[Ngo][14:40]

Yes, Wednesday would be good

[Yudkowsky][14:40]

let's call it tentatively for that

[Soares][14:41]

Great! Thanks for the chats.

[Ngo][14:41]

Thanks both!

[Yudkowsky][14:41]

Thanks, Richard!

2. Follow-ups

2.1. Richard Ngo's summary

[Tallinn][0:35] (Sep. 6)

just caught up here & wanted to thank nate, eliezer and (especially) richard for doing this! it's great to see eliezer's model being probed so intensively. i've learned a few new things (such as the genetic bottleneck being plausibly a big factor in human cognition). FWIW, a minor comment re deontology (as that's fresh on my mind): in my view deontology is more about coordination than optimisation: deontological agents are more trustworthy, as they're much easier to reason about (in the same way how functional/declarative code is easier to reason about than imperative code). hence my steelman of bureaucracies (as well as social norms): humans just (correctly) prefer their fellow optimisers (including non-human optimisers) to be deontological for trust/coordination reasons, and are happy to pay the resulting competence tax.

[Ngo][3:10] (Sep. 8)

Thanks Jaan! I agree that greater trust is a good reason to want agents which are deontological at some high level.

I've attempted a summary of the key points so far; comments welcome:
[GDocs link]

[Ngo] (Sep. 8 Google Doc)

1st discussion

(Mostly summaries not quotations)

Eliezer, summarized by Richard: "To avoid catastrophe, whoever builds AGI first will have to a) align it to some extent, and b) decide not to scale it up beyond the point where their alignment techniques fail, and c) do some pivotal act that prevents others from scaling it up to that level. But ~~our alignment techniques will not be good enough~~ ~~our alignment techniques will be very far from adequate~~ on our current trajectory, our alignment techniques will be very far from adequate to create an AI that safely performs any such pivotal act."

[Yudkowsky][11:05] (Sep. 8 comment)

will not be good enough

Are not presently on course to be good enough, missing by not a little.
"Will not be good enough" is literally declaring for lying down and dying.

[Yudkowsky][16:03] (Sep. 9 comment)

will [be very far from adequate]

Same problem as the last time I commented. I am not making an unconditional prediction about future failure as would be implied by the word "will". Conditional on current courses of action or their near neighboring courses, we seem to be well over an order of magnitude away from surviving, unless a miracle occurs. It's still in the end a result of people doing what they seem to be doing, not an inevitability.

[Ngo][5:10] (Sep. 10 comment)

Ah, I see. Does adding "on our current trajectory" fix this?

[Yudkowsky][10:46] (Sep. 10 comment)

Yes.

[Ngo] (Sep. 8 Google Doc)

Richard, summarized by Richard: "Consider the pivotal act of 'make a breakthrough in alignment research'. It is likely that, before the point where AGIs are strongly superhuman at seeking power, they will already be strongly superhuman at understanding the world, and at performing narrower pivotal acts like alignment research which don't require as much agency (by which I roughly mean: large-scale motivations and the ability to pursue them over long timeframes)."

Eliezer, summarized by Richard: "There's a deep connection between solving intellectual problems and taking over the world - the former requires a powerful mind to think about domains that, when solved, render very cognitively accessible strategies that can do dangerous things. Even mathematical research is a goal-oriented task which involves identifying then pursuing instrumental subgoals - and if brains which evolved to hunt on the savannah can quickly learn to do mathematics, then it's also plausible that AIs trained to do mathematics could quickly learn a range of other skills. Since almost nobody understands the deep similarities in the cognition required for these different tasks, the distance between AIs that are able to perform fundamental scientific research, and dangerously agentic AGIs, is smaller than almost anybody expects."

[Yudkowsky][11:05] (Sep. 8 comment)

There's a deep connection between solving intellectual problems and taking over the world

There's a deep connection by default between chipping flint handaxes and taking over the world, if you happen to learn how to chip handaxes in a very general way. "Intellectual" problems aren't special in this way. And maybe you could avert the default, but that would take some work and you'd have to do it before easier default ML techniques destroyed the world.

[Ngo] (Sep. 8 Google Doc)

Richard, summarized by Richard: "Our lack of understanding about how intelligence works also makes it easy to assume that traits which co-occur in humans will also co-occur in future AIs. But human brains are badly-optimised for tasks like scientific research, and well-optimised for seeking power over the world, for reasons including a) evolving while embodied in a harsh environment; b) the genetic bottleneck; c) social environments which rewarded power-seeking. By contrast, training neural networks on tasks like mathematical or scientific research optimises them much less for seeking power. For example, GPT-3 has knowledge and reasoning capabilities but little agency, and loses coherence when run for longer timeframes."

[Tallinn][4:19] (Sep. 8 comment)

[well-optimised for] seeking power

male-female differences might be a datapoint here (annoying as it is to lean on pinker's point :))

[Yudkowsky][11:31] (Sep. 8 comment)

I don't think a female Eliezer Yudkowsky doesn't try to save / optimize / takeover the world. Men may do that for nonsmart reasons; smart men and women follow the same reasoning when they are smart enough. Eg Anna Salamon and many others.

[Ngo] (Sep. 8 Google Doc)

Eliezer, summarized by Richard: "Firstly, there's a big difference between most scientific research and the sort of pivotal act that we're talking about - you need to explain how AIs with a given skill can be used to actually prevent dangerous AGIs from being built. Secondly, insofar as GPT-3 has little agency, that's because it has memorised many shallow patterns in a way which won't directly scale up to general intelligence. Intelligence instead consists of deep problem-solving patterns which link understanding and agency at a fundamental level."

3. September 8 conversation

3.1. The Brazilian university anecdote

[Yudkowsky][11:00]

(I am here.)

[Ngo][11:01]

Me too.

[Soares][11:01]

Welcome back!

(I'll mostly stay out of the way again.)

[Ngo][11:02]

Cool. Eliezer, did you read the summary - and if so, do you roughly endorse it?

Also, I've been thinking about the best way to approach discussing your intuitions about cognition. My guess is that starting with the obedience vs paperclips thread is likely to be less useful than starting somewhere else - e.g. the description you gave near the beginning of the last discussion, about "searching for states that get fed into a result function and then a result-scoring function".

[Yudkowsky][11:06]

made a couple of comments about phrasings in the doc

So, from my perspective, there's this thing where... it's really quite hard to teach certain *general* points by talking at people, as opposed to more specific points. Like, they're trying to build a perpetual motion machine, and even if you can manage to argue them into believing their first design is wrong, they go looking for a new design, and the new design is complicated enough that they can no longer be convinced that they're wrong because they managed to make a more complicated error whose refutation they couldn't keep track of anymore.

Teaching people to see an underlying structure in a lot of places is a very hard thing to teach in this way. Richard Feynman [gave an example](#) of the mental motion in his story that ends "Look at the water!", where people learned in classrooms about how "a medium with an index" is supposed to polarize light reflected from it, but they didn't realize that sunlight coming off of water would be polarized. My guess is that doing this properly requires homework exercises; and that, unfortunately from my own standpoint, it happens to be a place where I have extra math talent, the same way that eg Marcello is more talented at formally proving theorems than I happen to be; and that people without the extra math talent, have to do a lot *more* exercises than I did, and I don't have a good sense of which exercises to give them.

[Ngo][11:13]

I'm sympathetic to this, and can try to turn off skeptical-discussion-mode and turn on learning-mode, if you think that'll help.

[Yudkowsky][11:14]

There's a general insight you can have about how arithmetic is commutative, and for some people you can show them $1 + 2 = 2 + 1$ and their native insight suffices to generalize over the 1 and the 2 to any

other numbers you could put in there, and they realize that strings of numbers can be rearranged and all end up equivalent. For somebody else, when they're a kid, you might have to show them 2 apples and 1 apple being put on the table in a different order but ending up with the same number of apples, and then you might have to show them again with adding up bills in different denominations, in case they didn't generalize from apples to money. I can actually remember being a child young enough that I tried to add 3 to 5 by counting "5, 6, 7" and I thought there was some clever enough way to do that to actually get 7, if you tried hard.

Being able to see "consequentialism" is like that, from my perspective.

[Ngo][11:15]

Another possibility: can you trace the origins of this belief, and how it came out of your previous beliefs?

[Yudkowsky][11:15]

I don't know what homework exercises to give people to make them able to see "consequentialism" all over the place, instead of inventing slightly new forms of consequentialist cognition and going "Well, now *that* isn't consequentialism, right?"

Trying to say "searching for states that get fed into an input-result function and then a result-scoring function" was one attempt of mine to describe the dangerous thing in a way that would maybe sound abstract enough that people would try to generalize it more.

[Ngo][11:17]

Another possibility: can you describe the closest thing to real consequentialism in humans, and how it came about in us?

[Yudkowsky][11:18][11:21]

Ok, so, part of the problem is that... before you do enough homework exercises for whatever your level of talent is (and even I, at one point, had done little enough homework that I thought there might be a clever way to add 3 and 5 in order to get to 7), you tend to think that only the very crisp formal thing that's been presented to you, is the "real" thing.

Why would your engine have to obey the laws of thermodynamics? You're not building one of those Carnot engines you saw in the physics textbook!

Humans contain fragments of consequentialism, or bits and pieces whose interactions add up to partially imperfectly shadow consequentialism, and the critical thing is being able to see that the reason why humans' outputs 'work', in a sense, is because these structures are what is doing

the work, and the work gets done because of how they shadow consequentialism and only insofar as they shadow consequentialism.

Put a human in one environment, it gets food. Put a human in a different environment, it gets food again. Wow, different initial conditions, same output! There must be things inside the human that, whatever else they do, are also along the way somehow effectively searching for motor signals such that food is the end result!

[Ngo][11:20]

To me it feels like you're trying to nudge me (and by extension whoever reads this transcript) out of a specific failure mode. If I had to guess, something like: "I understand what Eliezer is talking about so now I'm justified in disagreeing with it", or perhaps "Eliezer's explanation didn't make sense to me and so I'm justified in thinking that his concepts don't make sense". Is that right?

[Yudkowsky][11:22]

More like... from my perspective, even after I talk people out of one specific perpetual motion machine being possible, they go off and try to invent a different, more complicated perpetual motion machine.

And I am not sure what to do about that. It has been going on for a very long time from my perspective.

In the end, a lot of what people got out of all that writing I did, was not the deep object-level principles I was trying to point to - they did not really get [Bayesianism as thermodynamics](#), say, they did not become able to see [Bayesian structures](#) any time somebody sees a thing and changes their belief. What they got instead was something much more meta and general, a vague spirit of how to reason and argue, because that was what they'd spent a lot of time being exposed to over and over and over again in lots of blog posts.

Maybe there's no way to make somebody understand why [corrigibility](#) is "unnatural" except to repeatedly walk them through the task of trying to invent an agent structure that lets you press the shutdown button (without it trying to force you to press the shutdown button), and showing them how each of their attempts fails; and then also walking them through why Stuart Russell's attempt at moral uncertainty produces the [problem of fully updated \(non-\)deference](#); and hope they can start to see the informal general pattern of why corrigibility is in general contrary to the structure of things that are good at optimization.

Except that to do the exercises at all, you need them to work within an expected utility framework. And then they just go, "Oh, well, I'll just build an agent that's good at optimizing things but doesn't use these explicit expected utilities that are the source of the problem!"

And then if I want them to believe the same things I do, for the same reasons I do, I would have to teach them why certain structures of

cognition are the parts of the agent that are good at stuff and do the work, rather than them being this particular formal thing that they learned for manipulating meaningless numbers as opposed to real-world apples.

And I have tried to write that page once or twice (eg "[coherent decisions imply consistent utilities](#)") but it has not sufficed to teach them, because they did not even do as many homework problems as I did, let alone the greater number they'd have to do because this is in fact a place where I have a particular talent.

I don't know how to solve this problem, which is why I'm falling back on talking about it at the meta-level.

[Ngo][11:30]

I'm reminded of a LW post called "[Write a thousand roads to Rome](#)", which iirc argues in favour of trying to explain the same thing from as many angles as possible in the hope that one of them will stick.

[Soares][11:31]

(Suggestion, not-necessarily-good: having named this problem on the meta-level, attempt to have the object-level debate, while flagging instances of this as it comes up.)

[Ngo][11:31]

I endorse Nate's suggestion.

And will try to keep the difficulty of the meta-level problem in mind and respond accordingly.

[Yudkowsky][11:33]

That (Nate's suggestion) is probably the correct thing to do. I name it out loud because sometimes being told about the meta-problem actually does help on the object problem. It seems to help me a lot and others somewhat less, but it does help others at all, for many others.

3.2. Brain functions and outcome pumps

[Yudkowsky][11:34]

So, do you have a particular question you would ask about input-seeking cognitions? I did try to say why I mentioned those at all (it's a different road to Rome on "consequentialism").

[Ngo][11:36]

Let's see. So the visual cortex is an example of quite impressive cognition in humans and many other animals. But I'd call this "pattern-recognition" rather than "searching for high-scoring results".

[Yudkowsky][11:37]

Yup! And it is no coincidence that there are no whole animals formed entirely out of nothing but a visual cortex!

[Ngo][11:37]

Okay, cool. So you'd agree that the visual cortex is doing something that's qualitatively quite different from the thing that animals overall are doing.

Then another question is: can you characterise searching for high-scoring results in non-human animals? Do they do it? Or are you mainly talking about humans and AGIs?

[Yudkowsky][11:39]

Also by the time you get to like the temporal lobes or something, there is probably some significant amount of "what could I be seeing that would produce this visual field?" that is searching through hypothesis-space for hypotheses with high plausibility scores, and for sure at the human level, humans will start to think, "Well, could I be seeing this? No, that theory has the following problem. How could I repair that theory?" But it is plausible that there is no low-level analogue of this in a monkey's temporal cortex; and even more plausible that the parts of the visual cortex, if any, which do anything analogous to this, are doing it in a relatively local and definitely very domain-specific way.

Oh, that's the cerebellum and motor cortex and so on, if we're talking about a cat or whatever. They have to find motor plans that result in their catching the mouse.

Just because the visual cortex isn't (obviously) running a search doesn't mean the rest of the animal isn't running any searches.

(On the meta-level, I notice myself hiccuping "But how could you not see that when looking at a cat?" and wondering what exercises would be required to teach that.)

[Ngo][11:41]

Well, I see *something* when I look at a cat, but I don't know how well it corresponds to the concepts you're using. So just taking it slowly for now.

I have the intuition, by the way, that the motor cortex is in some sense doing a similar thing to the visual cortex - just in reverse. So instead of taking low-level inputs and producing high-level outputs, it's taking high-level inputs and producing low-level outputs. Would you agree with that?

[Yudkowsky][11:43]

It doesn't directly parse in my ontology because (a) I don't know what you mean by 'high-level' and (b) whole Cartesian agents can be viewed as functions, that doesn't mean all agents can be viewed as non-searching pattern-recognizers.

That said, all parts of the cerebral cortex have surprisingly similar morphology, so it wouldn't be at all surprising if the motor cortex is doing something similar to visual cortex. (The cerebellum, on the other hand...)

[Ngo][11:44]

The signal from the visual cortex saying "that is a cat", and the signal to the motor cortex saying "grab that cup", are things I'd characterise as high-level.

[Yudkowsky][11:45]

Still less of a native distinction in my ontology, but there's an informal thing it can sort of wave at, and I can hopefully take that as understood and run with it.

[Ngo][11:45]

The firing of cells in the retina, and firing of motor neurons, are the low-level parts.

Cool. So to a first approximation, we can think about the part in between the cat recognising a mouse, and the cat's motor cortex producing the specific neural signals required to catch the mouse, as the part where the consequentialism happens?

[Yudkowsky][11:49]

The part between the cat's eyes seeing the mouse, and the part where the cat's limbs move to catch the mouse, is the whole cat-agent. The

whole cat agent sure is a baby consequentialist / searches for mouse-catching motor patterns / gets similarly high-scoring end results even as you vary the environment.

The visual cortex is a particular part of this system-viewed-as-a-feedforward-function that is, plausibly, by no means surely, either not very searchy, or does only small local visual-domain-specific searches not aimed per se at catching mice; it has the epistemic nature rather than the planning nature.

Then from one perspective you could reason that "well, most of the consequentialism is in the remaining cat after visual cortex has sent signals onward". And this is in general a dangerous mode of reasoning that is liable to fail in, say, inspecting every particular neuron for consequentialism and not finding it; but in this particular case, there are significantly more consequentialist parts of the cat than the visual cortex, so I am okay running with it.

[Ngo][11:50]

Ah, the more specific thing I meant to say is: most of the consequentialism is strictly between the visual cortex and the motor cortex. Agree/disagree?

[Yudkowsky][11:51]

Disagree, I'm rusty on my neuroanatomy but I think the motor cortex may send signals on to the cerebellum rather than the other way around.

(I may also disagree with the actual underlying notion you're trying to hint at, so possibly not just a "well include the cerebellum then" issue, but I think I should let you respond first.)

[Ngo][11:53]

I don't know enough neuroanatomy to chase that up, so I was going to try a different tack.

But actually, maybe it's easier for me to say "let's include the cerebellum" and see where you think the disagreement ends up.

[Yudkowsky][11:56]

So since cats are not (obviously) (that I have read about) cross-domain consequentialists with imaginations, their consequentialism is in bits and pieces of consequentialism embedded in them all over by the more purely pseudo-consequentialist genetic optimization loop that built them.

A cat who fails to catch a mouse may then get little bits and pieces of catbrain adjusted all over.

And then those adjusted bits and pieces get a pattern lookup later.

Why do these pattern-lookups with no obvious immediate search element, all happen to point towards the same direction of catching the mouse? Because of the past causal history about how what gets looked up, which was tweaked to catch the mouse.

So it is legit harder to point out "the consequentialist parts of the cat" by looking for which sections of neurology are doing searches right there. That said, to the extent that the visual cortex does not get tweaked on failure to catch a mouse, it's not part of that consequentialist loop either.

And yes, the same applies to humans, but humans also do more explicitly searchy things and this is part of the story for why humans have spaceships and cats do not.

[Ngo][12:00]

Okay, this is interesting. So in biological agents we've got these three levels of consequentialism: evolution, reinforcement learning, and planning.

[Yudkowsky][12:01]

In biological agents we've got evolution + local evolved system-rules that in the past promoted genetic fitness. Two kinds of local rules like this are "operant-conditioning updates from success or failure" and "search through visualized plans". I wouldn't characterize these two kinds of rules as "levels".

[Ngo][12:02]

Okay, I see. And when you talk about searching through visualised plans (the type of thing that humans do) can you say more about what it means for that to be a "search"?

For example, if I imagine writing a poem line-by-line, I may only be planning a few words ahead. But somehow the whole poem, which might be quite long, ends up a highly-optimised product. Is that a central example of planning?

[Yudkowsky][12:04][12:07]

Planning is one way to succeed at search. I think for purposes of understanding alignment difficulty, you want to be thinking on the level of abstraction where you see that in some sense it is the search itself that is dangerous when it's a strong enough search, rather than the danger seeming to come from details of the planning process.

One of my early experiences in successfully generalizing my notion of intelligence, what I'd later verbalize as "computationally efficient finding of actions that produce outcomes high in a preference ordering", was in writing an (unpublished) story about time-travel in which the universe was globally consistent.

The requirement of global consistency, the way in which all events between Paradox start and Paradox finish had to map the Paradox's initial conditions onto the endpoint that would go back and produce those exact initial conditions, ended up imposing strong complicated constraints on reality that the Paradox in effect had to navigate using its initial conditions. The time-traveler needed to end up going through certain particular experiences that would produce the state of mind in which he'd take the actions that would end up prodding his future self elsewhere into having those experiences.

The Paradox ended up killing the people who built the time machine, for example, because they would not otherwise have allowed that person to go back in time, or kept the temporal loop open that long for any other reason if they were still alive.

Just having two examples of strongly consequentialist general optimization in front of me - human intelligence, and evolutionary biology - hadn't been enough for me to properly generalize over a notion of optimization. Having three examples of homework problems I'd worked - human intelligence, evolutionary biology, and the fictional Paradox - caused it to finally click for me.

[Ngo][12:07]

Hmm. So to me, one of the central features of search is that you consider many possibilities. But in this poem example, I may only have explicitly considered a couple of possibilities, because I was only looking ahead a few words at a time. This seems related to the distinction Abram drew a while back between selection and control (<https://www.alignmentforum.org/posts/ZDZmopKquzHYPRNxq/selection-vs-control>). Do you distinguish between them in the same way as he does? Or does "control" of a system (e.g. a football player dribbling a ball down the field) count as search too in your ontology?

[Yudkowsky][12:10][12:11]

I would later try to tell people to "imagine a paperclip maximizer as *not being a mind at all*, imagine it as a kind of malfunctioning time machine that spits out outputs which will in fact result in larger numbers of paperclips coming to exist later". I don't think it clicked because people hadn't done the same homework problems I had, and didn't have the same "Aha!" of realizing how part of the notion and danger of intelligence could be seen in such purely material terms.

But the [convergent instrumental strategies](#), the anticorrigibility, these things are contained in the *true fact about the universe* that certain outputs of the time machine *will in fact* result in there being lots more paperclips later. What produces the danger is not the details of the search process, it's the search being strong and effective *at all*. The danger is in the territory itself and not just in some weird map of it; that building nanomachines that kill the programmers will produce more paperclips is a fact about reality, not a fact about paperclip maximizers!

[Ngo][12:11]

Right, I remember a very similar idea in your writing about Outcome Pumps (<https://www.lesswrong.com/posts/4ARaTpNX62ual86j6/the-hidden-complexity-of-wishes>).

[Yudkowsky][12:12]

Yup! Alas, the story was written in 2002-2003 when I was a worse writer and the real story that inspired the Outcome Pump never did get published.

[Ngo][12:14]

Okay, so I guess the natural next question is: what is it that makes you think that a strong, effective search isn't likely to be limited or constrained in some way?

What is it about search processes (like human brains) that makes it hard to train them with blind spots, or deontological overrides, or things like that?

Hmmm, although it feels like this is a question I can probably predict your answer to. (Or maybe not, I wasn't expecting the time travel.)

[Yudkowsky][12:15]

In one sense, they are! A paperclip-maximizing superintelligence is nowhere near as powerful as a paperclip-maximizing time machine. The time machine can do the equivalent of buying winning lottery tickets from lottery machines that have been thermodynamically randomized; a superintelligence can't, at least not directly without rigging the lottery or whatever.

But a paperclip-maximizing strong general superintelligence is epistemically and instrumentally [efficient](#), relative to *you*, or to me. Any time we see it can get at least X paperclips by doing Y, we should expect that it gets X or more paperclips by doing Y or something that leads to even more paperclips than that, because it's not going to miss the strategy we see.

So in that sense, searching our own brains for how a time machine would get paperclips, asking ourselves how many paperclips are in principle possible and how they could be obtained, is a way of getting our own brains to consider lower bounds on the problem without the implicit stupidity assertions that our brains unwittingly use to constrain story characters. Part of the point of telling people to think about time machines instead of superintelligences was to get past the ways they imagine superintelligences being stupid. Of course that didn't work either, but it was worth a try.

I don't think that's quite what you were asking about, but I want to give you a chance to see if you want to rephrase anything before I try to answer your me-reformulated questions.

[Ngo][12:20]

Yeah, I think what I wanted to ask is more like: why should we expect that, out of the space of possible minds produced by optimisation algorithms like gradient descent, strong general superintelligences are more common than other types of agents which score highly on our loss functions?

[Yudkowsky][12:20][12:23][12:24]

It depends on how hard you optimize! And whether gradient descent on a particular system can even successfully optimize that hard! Many current AIs are trained by gradient descent and yet not superintelligences at all.

But the answer is that some problems are difficult in that they require solving lots of subproblems, and an easy way to solve all those subproblems is to use patterns which collectively have some coherence and overlap, and the coherence within them generalizes across all the subproblems. Lots of search orderings will stumble across something like that before they stumble across separate solutions for lots of different problems.

I suspect that you cannot get this out of small large amounts of gradient descent on small large layered transformers, and therefore I suspect that GPT-N does not approach superintelligence before the world is ended by systems that look differently, but I could be wrong about that.

[Ngo][12:22][12:23]

Suppose that we optimise hard enough to produce an epistemic subsystem that can make plans much better than any human's.

My guess is that you'd say that this is *possible*, but that we're much more likely to first produce a consequentialist agent which does this (rather than a purely epistemic agent which does this).

[Yudkowsky][12:24]

I am confused by what you think it means to have an "epistemic subsystem" that "makes plans much better than any human's". If it searches paths through time and selects high-scoring ones for output, what makes it "epistemic"?

[Ngo][12:25]

Suppose, for instance, that it doesn't actually carry out the plans, it just writes them down for humans to look at.

[Yudkowsky][12:25]

If it *can in fact* do the thing that a paperclipping time machine does, what makes it any safer than a paperclipping time machine because we called it "epistemic" or by some other such name?

By what criterion is it selecting the plans that humans look at?

Why did it make a difference that its output was fed through the causal systems called humans on the way to the causal systems called protein synthesizers or the Internet or whatever? If we build a superintelligence to design nanomachines, it makes no obvious difference to its safety whether it sends DNA strings directly to a protein synthesis lab, or humans read the output and retype it manually into an email. Presumably you also don't think that's where the safety difference comes from. So where does the safety difference come from?

(note: lunchtime for me in 2 minutes, propose to reconvene in 30m after that)

[Ngo][12:28]

(break for half an hour sounds good)

If we consider the visual cortex at a given point in time, how does it decide which objects to recognise?

Insofar as the visual cortex can be non-consequentialist about which objects it recognises, why couldn't a planning system be non-consequentialist about which plans it outputs?

[Yudkowsky][12:32]

This does feel to me like another "look at the water" moment, so what do you predict I'll say about that?

[Ngo][12:34]

I predict that you say something like: in order to produce an agent that can create very good plans, we need to apply a lot of optimisation power to that agent. And if the channel through which we're applying that optimisation power is "giving feedback on its plans", then we don't have a mechanism to ensure that the agent actually learns to optimise for creating really good plans, as opposed to creating plans that receive really good feedback.

[Soares][12:35]

Seems like a fine cliffhanger?

[Ngo][12:35]

Yepp.

[Soares][12:35]

Great. Let's plan to reconvene in 30min.

3.3. Hypothetical-planning systems, nanosystems, and evolving generality

[Yudkowsky][13:03][13:11]

So the answer you expected from me, translated into my terms, would be, "If you select for the consequence of the humans hitting 'approve' on the plan, you're still navigating the space of inputs for paths through time to probable outcomes (namely the humans hitting 'approve'), so you're still doing consequentialism."

But suppose you manage to avoid that. Suppose you get exactly what you ask for. Then the system is still outputting *plans* such that, when humans follow them, they take paths through time and end up with outcomes that score high in some scoring function.

My answer is, "What the heck would it mean for a *planning system* to be *non-consequentialist*? You're asking for nonwet water! What's consequentialist isn't the system that does the work, it's the work you're trying to do! You could imagine it being done by a cognition-free material system like a time machine and it would still be consequentialist because the output is a *plan*, a path through time!"

And this indeed is a case where I feel a helpless sense of not knowing how I can rephrase things, which exercises you have to get somebody to do, what fictional experience you have to walk somebody through, before they start to look at the water and see a material with an index, before they start to look at the phrase "why couldn't a planning system be non-consequentialist about which plans it outputs" and go "um".

My imaginary listener now replies, "Ah, but what if we have plans that don't end up with outcomes that score high in some function?" and I reply "Then you lie on the ground randomly twitching because any *outcome you end up with* which is *not that* is one that you wanted *more than that* meaning you *preferred it more than the outcome of random motor outputs* which is *optimization toward higher in the preference function* which is *taking a path through time that leads to particular destinations more than it leads to random noise*."

[Ngo][13:09][13:11]

Yeah, this does seem like a good example of the thing you were trying to explain at the beginning

It still feels like there's some sort of levels distinction going on here though, let me try to tease out that intuition.

Okay, so suppose I have a planning system that, given a situation and a goal, outputs a plan that leads from that situation to that goal.

And then suppose that we give it, as input, a situation that we're not actually in, and it outputs a corresponding plan.

It seems to me that there's a difference between the sense in which that planning system is consequentialist by virtue of making consequentialist plans (as in: if that plan were used in the situation described in its inputs, it would lead to some goal being achieved) versus another hypothetical agent that is just directly trying to achieve goals in the situation it's actually in.

[Yudkowsky][13:18]

So I'd preface by saying that, *if* you could build such a system, which is indeed a coherent thing (it seems to me) to describe for the purpose of building it, then there would possibly be a safety difference on the margins, it would be noticeably less dangerous though still dangerous. It would need a special internal structural property that you might not get by gradient descent on a loss function with that structure, just like natural selection on inclusive genetic fitness doesn't get you explicit fitness optimizers; you could optimize for planning in hypothetical situations, and get something that didn't explicitly care only and strictly about hypothetical situations. And even if you did get that, the outputs that would kill or brain-corrupt the operators in hypothetical situations might also be fatal to the operators in actual situations. But that is a coherent

thing to describe, and the fact that it was not optimizing our own universe, might make it *safer*.

With that said, I would worry that somebody would think there was some bone-deep difference of agentiness, of something they were empathizing with like personhood, of imagining goals and drives being absent or present in one case or the other, when they imagine a planner that just solves "hypothetical" problems. If you take that planner and feed it the actual world as its hypothetical, tada, it is now that big old dangerous consequentialist you were imagining before, without it having acquired some difference of *psychological* agency or 'caring' or whatever.

So I think there is an important homework exercise to do here, which is something like, "Imagine that safe-seeming system which only considers hypothetical problems. Now see that if you take that system, don't make any other internal changes, and feed it actual problems, it's very dangerous. Now meditate on this until you can see how the hypothetical-considering planner was extremely close in the design space to the more dangerous version, had all the dangerous latent properties, and would probably have a bunch of actual dangers too."

"See, you thought the source of the danger was this internal property of caring about actual reality, but it wasn't that, it was the structure of planning!"

[Ngo][13:22]

I think we're getting closer to the same page now.

Let's consider this hypothetical planner for a bit. Suppose that it was trained in a way that minimised the, let's say, *adversarial* component of its plans.

For example, let's say that the plans it outputs for any situation are heavily regularised so only the broad details get through.

Hmm, I'm having a bit of trouble describing this, but basically I have an intuition that in this scenario there's a component of its plan which is cooperative with whoever executes the plan, and a component that's adversarial.

And I agree that there's no fundamental difference in type between these two things.

[Yudkowsky][13:27]

"What if this potion we're brewing has a Good Part and a Bad Part, and we could just keep the Good Parts..."

[Ngo][13:27]

Nor do I think they're separable. But in some cases, you might expect one to be much larger than the other.

[Soares][13:29]

(I observe that my model of some other listeners, at this point, protest "there is yet a difference between the hypothetical-planner applied to actual problems, and the Big Scary Consequentialist, which is that the hypothetical planner is emitting descriptions of plans that *would* work if executed, whereas the big scary consequentialist is executing those plans directly.")

(Not sure that's a useful point to discuss, or if it helps Richard articulate, but it's at least a place I expect some reader's minds to go if/when this is published.)

[Yudkowsky][13:30]

(That is in fact a difference! The insight is in realizing that the hypothetical planner is only one line of outer shell command away from being a Big Scary Thing and is therefore also liable to be Big and Scary in many ways.)

[Ngo][13:31]

To me it seems that Eliezer's position is something like: "actually, in almost no training regimes do we get agents that decide which plans to output by spending almost all of their time thinking about the object-level problem, and very little of their time thinking about how to manipulate the humans carrying out the plan".

[Yudkowsky][13:32]

My position is that the AI does not neatly separate its internals into a Part You Think Of As Good and a Part You Think Of As Bad, because that distinction is sharp in your map but not sharp in the territory or the AI's map.

From the perspective of a paperclip-maximizing-action-outputting-time-machine, its actions are not "object-level making paperclips" or "manipulating the humans next to the time machine to deceive them about what the machine does", they're just physical outputs that go through time and end up with paperclips.

[Ngo][13:34]

@Nate, yeah, that's a nice way of phrasing one point I was trying to make. And I do agree with Eliezer that these things *can be* very similar.

But I'm claiming that in some cases these things can also be quite different - for instance, when we're training agents that only get to output a short high-level description of the plan.

[Yudkowsky][13:35]

The danger is in how hard the agent has to work to come up with the plan. I can, for instance, build an agent that very safely outputs a high-level plan for saving the world:

echo "Hey Richard, go save the world!"

So I do have to ask what kind of "high-level" planning output, that saves the world, you are envisioning, and why it was hard to cognitively come up with such that we didn't just make that high-level plan right now, if humans could follow it. Then I'll look at the part where the plan was hard to come up with, and say how the agent had to understand lots of complicated things in reality and accurately navigate paths through time for those complicated things, in order to even invent the high-level plan, and hence it was very dangerous if it wasn't navigating exactly where you hoped. Or, alternatively, I'll say, "That plan couldn't save the world: you're not postulating enough superintelligence to be dangerous, and you're also not using enough superintelligence to flip the tables on the currently extremely doomed world."

[Ngo][13:39]

At this point I'm not envisaging a particular planning output that saves the world, I'm just trying to get more clarity on the issue of consequentialism.

[Yudkowsky][13:40]

Look at the water; it's not the way you're doing the work that's dangerous, it's the work you're trying to do. What work are you trying to do, never mind how it gets done?

[Ngo][13:41]

I think I agree with you that, in the limit of advanced capabilities, we can't say much about how the work is being done, we have to primarily reason from the work that we're trying to do.

But here I'm only talking about systems that are intelligent enough to come up with plans and do research that are beyond the capability of humanity.

And for me the question is: for *those* systems, can we tilt the way they do the work so they spend 99% of their time trying to solve the object-level problem, and 1% of their time trying to manipulate the humans who are

going to carry out the plan? (Where these are not fundamental categories for the AI, they're just a rough categorisation that emerges after we've trained it - the same way that the categories of "physically moving around" and "thinking about things" aren't fundamentally different categories of action for humans, but the way we've evolved means there's a significant internal split between them.)

[Soares][13:43]

(I suspect Eliezer is not trying to make a claim of the form "in the limit of advanced capabilities, we are relegated to reasoning about what work gets done, not about how it was done". I suspect some miscommunication. It might be a reasonable time for Richard to attempt to paraphrase Eliezer's argument?)

(Though it also seems to me like Eliezer responding to the 99%/1% point may help shed light.)

[Yudkowsky][13:46]

Well, for one thing, I'd note that a system which is designing nanosystems, and spending 1% of its time thinking about how to kill the operators, is lethal. It has to be such a small fraction of thinking that it, like, never completes the whole thought about "well, if I did X, that would kill the operators!"

[Ngo][13:46]

Thanks for that, Nate. I'll try to paraphrase Eliezer's argument now.

Eliezer's position (partly in my own terminology): we're going to build AIs that can perform very difficult tasks using cognition which we can roughly describe as "searching over many options to find one that meets our criteria". An AI that can solve these difficult tasks will need to be able to search in a very general and flexible way, and so it will be very difficult to constrain that search into a particular region.

Hmm, that felt like a very generic summary, let me try and think about the more specific claims he's making.

[Yudkowsky][13:54]

An AI that can solve these difficult tasks will need to be able to

Very very little is universally necessary over the design space. The *first* AGI that our tech becomes able to build is liable to work in certain easier and simpler ways.

[Ngo][13:55]

Point taken; thanks for catching this misphrasing (this and previous times).

[Yudkowsky][13:56]

Can you, in principle, build a red-car-driver that is totally incapable of driving blue cars? In principle, sure! But the first red-car-driver that gradient descent stumbles over is liable to be a blue-car-driver too.

[Ngo][13:57]

Eliezer, I'm wondering how much of our disagreement is about how high the human level is here.

Or, to put it another way: we can build systems that outperform humans at quite a few tasks by now, without having search abilities that are general enough to even try to take over the world.

[Yudkowsky][13:58]

Indubitably and indeed, this is so.

[Ngo][13:59]

Putting aside for a moment the question of which tasks are pivotal enough to save the world, which parts of your model draw the line between human-level chess players and human-level galaxy-colonisers?

And say that we'll be able to align ones that they outperform us on *these tasks* before taking over the world, but not on *these other tasks*?

[Yudkowsky][13:59][14:01]

That doesn't have a very simple answer, but one aspect there is *domain generality* which in turn is achieved through *novel domain learning*.

Humans, you will note, were not aggressively optimized by natural selection to be able to breathe underwater or fly into space. In terms of obvious outer criteria, there is not much outer sign that natural selection produced these creatures much more general than chimpanzees, by training on a much wider range of environments and loss functions.

[Soares][14:00]

(Before we drift too far from it: thanks for the summary! It seemed good to me, and I updated towards the miscommunication I feared not-having-happened.)

[Ngo][14:03]

(Before we drift too far from it: thanks for the summary! It seemed good to me, and I updated towards the miscommunication I feared not-having-happened.)

(Good to know, thanks for keeping an eye out. To be clear, I didn't ever interpret Eliezer as making a claim explicitly about the limit of advanced capabilities; instead it just seemed to me that he was thinking about AIs significantly more advanced than the ones I've been thinking of. I think I phrased my point poorly.)

[Yudkowsky][14:05][14:10]

There are complicated aspects of this story where natural selection may metaphorically be said to have "had no idea of what it was doing", eg, after early rises in intelligence possibly produced by sexual selection on neatly chipped flint handaxes or whatever, all the cumulative brain-optimization on chimpanzees reached a point where there was suddenly a sharp selection gradient on relative intelligence at Machiavellian planning against other humans (even more so than in the chimp domain) as a subtask of inclusive genetic fitness, and so continuing to optimize on "inclusive genetic fitness" in the same old savannah, turned out to happen to be optimizing hard on the subtask and internal capability of "outwit other humans", which optimized hard on "model other humans", which was a capability that could be reused for modeling the chimp-that-is-this-chimp, which turned the system on itself and made it reflective, which contributed greatly to its intelligence being generalized, even though it was just grinding the same loss function on the same savannah; the system being optimized happened to go there in the course of being optimized even harder for the same thing.

So one can imagine asking the question: Is there a superintelligent AGI that can quickly build nanotech, which has a kind of passive safety in some if not all respects, in virtue of it solving problems like "build a nanotech system which does X" the way that a beaver solves building dams, in virtue of having a bunch of specialized learning abilities without it ever having a cross-domain general learning ability?

And in this regard one does note that there are many, many, many things that humans do which no other animal does, which you might think would contribute a lot to that animal's fitness if there were animalistic ways to do it. They don't make iron claws for themselves. They never did evolve a tendency to search for iron ore, and burn wood into charcoal that could be used in hardened-clay furnaces.

No animal plays chess, but AIs do, so we can obviously make AIs to do things that animals don't do. On the other hand, the environment didn't exactly present any particular species with a challenge of chess-playing either.

Even so, though, even if some animal had evolved to play chess, I fully expect that current AI systems would be able to squish it at chess,

because the AI systems are on chips that run faster than neurons and doing crisp calculations and there are things you just can't do with noisy slow neurons. So that again is not a generally reliable argument about what AIs can do.

[Ngo][14:09][14:11]

Yes, although I note that challenges which are trivial from a human-engineering perspective can be very challenging from an evolutionary perspective (e.g. spinning wheels).

And so the evolution of animals-with-a-little-bit-of-help-from-humans might end up in very different places from the evolution of animals-just-by-themselves. And analogously, the ability of humans to fill in the gaps to help less general AIs achieve more might be quite significant.

[Yudkowsky][14:11]

So we can again ask: Is there a way to make an AI system that is *only* good at designing nanosystems, which can achieve some complicated but hopefully-specifiable real-world outcomes, without that AI also being superhuman at understanding and manipulating humans?

And I roughly answer, "Perhaps, but not by default, there's a bunch of subproblems, I don't actually know how to do it right now, it's not *the easiest* way to get an AGI that can build nanotech (and kill you), you've got to make the red-car-driver specifically not be able to drive blue cars." Can I explain how I know that? I'm really not sure I can, in real life where I explain X₀ and then the listener doesn't generalize X₀ to X and respecialize it to X₁.

It's like asking me how I could possibly know in 2008, before anybody had observed AlphaFold 2, that superintelligences would be able to crack the protein folding problem on the way to nanotech, which some people did question back in 2008.

Though that was admittedly more of a slam-dunk than this was, and I could not have told you that AlphaFold 2 would become possible at a prehuman level of general intelligence in 2021 specifically, or that it would be synced in time to a couple of years after GPT-2's level of generality at text.

[Ngo][14:18]

What are the most relevant axes of difference between solving protein folding and designing nanotech that, say, self-assembles into a computer?

[Yudkowsky][14:20]

Definitely, "turns out it's easier than you thought to use gradient descent's memorization of zillions of shallow patterns that overlap and recombine into larger cognitive structures, to add up to a consequentialist nanoengineer that only does nanosystems and never does sufficiently general learning to apprehend the big picture containing humans, while still understanding the goal for that pivotal act you wanted to do" is among the more plausible advance-specified miracles we could get.

But it is not what my model says actually happens, and I am not a believer that when your model says you are going to die, you get to start believing in particular miracles. You need to hold your mind open for any miracle and a miracle you didn't expect or think of in advance, because at this point our last hope is that in fact the future is often quite surprising - though, alas, negative surprises are a tad more frequent than positive ones, when you are trying desperately to navigate using a bad map.

[Ngo][14:22]

Perhaps one metric we could use here is something like: how much extra reward does the consequentialist nanoengineer get from starting to model humans, versus from becoming better at nanoengineering?

[Yudkowsky][14:23]

But that's *not* where humans came from. We didn't get to nuclear power by getting a bunch of fitness from nuclear power plants. We got to nuclear power because if you get a bunch of fitness from chipping flint handaxes and Machiavellian scheming, as found by relatively simple and local hill-climbing, that entrains the same genes that build nuclear power plants.

[Ngo][14:24]

Only in the specific case where you also have the constraint that you keep having to learn new goals every generation.

[Yudkowsky][14:24]

Huh???

[Soares][14:24]

(I think Richard's saying, "that's a consequence of the genetic bottleneck")

[Ngo][14:25]

Right.

Hmm, but I feel like we may have covered this ground before.

Suggestion: I have a couple of other directions I'd like to poke at, and then we could wrap up in 20 or 30 minutes?

[Yudkowsky][14:27]

OK

What are the most relevant axes of difference between solving protein folding and designing nanotech that, say, self-assembles into a computer?

Though I want to mark that this question seemed potentially cruxy to me, though perhaps not for others. I.e., if building protein factories that built nanofactories that built nanomachines that met a certain deep and lofty engineering goal, didn't involve cognitive challenges different in kind from protein folding, we could maybe just safely go do that using AlphaFold 3, which would be just as safe as AlphaFold 2.

I don't think we can do that. And I would note to the generic Other that if, to them, these both just sound like thinky things, so why can't you just do that other thinky thing too using the thinky program, this is a case where having any specific model of why we don't already have this nanoengineer right now would tell you there were specific different thinky things involved.

3.4. Coherence and pivotal acts

[Ngo][14:31]

In either order:

- I'm curious how the things we've been talking about relate to your opinions about meta-level optimisation from the AI foom debate. (I.e. talking about how wrapping around so that there's no longer any protected level of optimisation leads to dramatic change.)
- I'm curious how your claims about the "robustness" of consequentialism (i.e. the difficulty of channeling an agent's thinking in the directions we want it to go) relate to the reliance of humans on culture, and in particular the way in which humans raised without culture are such bad consequentialists.

On the first: if I were to simplify to the extreme, it seems like there are these two core intuitions that you've been trying to share for a long time. One is a certain type of recursive improvement, and another is a certain type of consequentialism.

[Yudkowsky][14:32]

The second question didn't make much sense in my native ontology? Humans raised without culture don't have access to environmental constants whose presence their genes assume, so they end up as broken machines and then they're bad consequentialists.

[Ngo][14:35]

Hmm, good point. Okay, question modification: the ways in which humans reason, act, etc, vary greatly depending on which cultures they're raised in. (I'm mostly thinking about differences over time - e.g. cavemen vs moderns.) My low-fidelity version of your view about consequentialists says that general consequentialists like humans possess a robust search process which isn't so easily modified.

(Sorry if this doesn't make much sense in your ontology, I'm getting a bit tired.)

[Yudkowsky][14:36]

What is it that varies that you think I think should predict would stay more constant?

[Ngo][14:37]

Goals, styles of reasoning, deontological constraints, level of conformity.

[Yudkowsky][14:39]

With regards to your first point, my first reaction was, "I just have one view of intelligence, what you see me arguing about reflects which points people have proved weirdly obstinate about. In 2008, Robin Hanson was being weirdly obstinate about how capabilities scaled and whether there was even any point in analyzing AIs differently from ems, so I talked about what I saw as the most slam-dunk case for there being Plenty Of Room Above Biology and for stuff going whoosh once it got above the human level.

"It later turned out that capabilities started scaling a whole lot *without* self-improvement, which is an example of the kind of weird surprise the Future throws at you, and maybe a case where I missed something by arguing with Hanson instead of imagining how I could be wrong in either

direction and not just the direction that other people wanted to argue with me about.

"Later on, people were unable to understand why alignment is hard, and got stuck on generalizing the concept I refer to as consequentialism. A theory of why I talked about both things for related reasons would just be a theory of why people got stuck on these two points for related reasons, and I think that theory would mainly be overexplaining an accident because if Yann LeCun had been running effective altruism I would have been explaining different things instead, after the people who talked a lot to EAs got stuck on a different point."

Returning to your second point, humans are broken things; if it were possible to build computers while working even worse than humans, we'd be having this conversation at that level of intelligence instead.

[Ngo][14:41]

(Retracted) I entirely agree about humans, but it doesn't matter that much how broken humans are when the regime of AIs that we're talking about is the regime that's directly above humans, and therefore only a bit less broken than humans.

[Yudkowsky][14:41]

Among the things to bear in mind about that, is that we then get tons of weird phenomena that are specific to humans, and you may be very out of luck if you start wishing for the *same* weird phenomena in AIs. Yes, even if you make some sort of attempt to train it using a loss function.

However, it does seem to me like as we start getting towards the Einstein level instead of the village-idiot level, even though this is usually not much of a difference, we do start to see the atmosphere start to thin already, and the turbulence start to settle down already. Von Neumann was actually a fairly reflective fellow who knew about, and indeed helped generalize, utility functions. The great achievements of von Neumann were not achieved by some very specialized hypernerd who spent all his fluid intelligence on crystallizing math and science and engineering alone, and so never developed any opinions about politics or started thinking about whether or not he had a utility function.

[Ngo][14:44]

I don't think I'm asking for the *same* weird phenomena. But insofar as a bunch of the phenomena I've been talking about have seemed weird according to your account of consequentialism, then the fact that approximately-human-level-consequentialists have lots of weird things about them is a sign that the phenomena I've been talking about are less unlikely than you expect.

[Yudkowsky][14:45][14:46]

I suspect that some of the difference here is that I think you have to be *noticeably* better than a human at nanoengineering to pull off pivotal acts large enough to make a difference, which is why I am not instead trying to gather the smartest people left alive and doing that pivotal act directly.

I can't think of anything you can do with somebody just barely smarter than a human, which flips the gameboard, aside of course from "go build a Friendly AI" which I *did* try to set up to just go do and which would be incredibly hard to align if we wanted an AI to do it instead (full-blown chicken-and-egg, that AI is already fully aligned).

[Ngo][14:45]

Oh, interesting. Actually one more question then: to what extent do you think that explicitly reasoning about utility functions and laws of rationality is what makes consequentialists have the properties you've been talking about?

[Yudkowsky][14:47, moved up in log]

Explicit reflection is one possible later stage of the path; an earlier part of the path is from being optimized to do things difficult enough that you need to stop stepping on your own feet and have different parts of your thoughts work well together.

It's the sort of path that has only one destination at its end, so there will be many ways to get there.

(Modulo various cases where different decision theories seem reflectively consistent and so on; I want to say "you know what I mean" but maybe people don't.)

[Ngo][14:47, moved down in log]

I suspect that some of the difference here is that I think you have to be *noticeably* better than a human at nanoengineering to pull off pivotal acts large enough to make a difference, which is why I am not instead trying to gather the smartest people left alive and doing that pivotal act directly.

Yepp, I think there's probably some disagreements about geopolitics driving this too. E.g. in my earlier summary document I mentioned some possible pivotal acts:

- Monitoring all potential AGI projects to an extent that makes it plausible for the US and China to work on a joint project without worrying that the other is privately racing.
- Provide arguments/demonstrations/proofs related to impending existential risk that are sufficiently compelling to scare the key global decision-makers into bottlenecking progress.

I predict that you think these would not be pivotal enough; but I don't think digging into the geopolitical side of things is the best use of our time.

[Yudkowsky][14:49, moved up in log]

Monitoring all AGI projects - either not politically feasible in real life given the actual way that countries behave in history books instead of fantasy; or at politically feasible levels, does not work well enough to prevent the world from ending once the know-how proliferates. The AI isn't doing much work here either; why not go do this now, if it's possible? (Note: please don't try to go do this now, it backfires badly.)

Provide sufficiently compelling arguments = superhuman manipulation, an incredibly dangerous domain that is just about the worst domain to try to align.

[Ngo][14:49, moved down in log]

With regards to your first point, my first reaction was, "I just have one view of intelligence, what you see me arguing about reflects which points people have proved weirdly obstinate about. In 2008, Robin Hanson was being weirdly obstinate about how capabilities scaled and whether there was even any point in analyzing AIs differently from ems, so I talked about what I saw as the most slam-dunk case for there being Plenty Of Room Above Biology and for stuff going whoosh once it got above the human level."

"It later turned out that capabilities started scaling a whole lot *without* self-improvement, which is an example of the kind of weird surprise the Future throws at you, and maybe a case where I missed something by arguing with Hanson instead of imagining how I could be wrong in either direction and not just the direction that other people wanted to argue with me about."

"Later on, people were unable to understand why alignment is hard, and got stuck on generalizing the concept I refer to as consequentialism. A theory of why I talked about both things for related reasons would just be a theory of why people got stuck on these two points for related reasons, and I think that theory would mainly be overexplaining an accident because if Yann LeCun had been running effective altruism I would have been explaining different things instead, after the people who talked a lot to EAs got stuck on a different point."

On my first point, it seems to me that your claims about recursive self-improvement were off in a fairly similar way to how I think your claims about consequentialism are off - which is that they defer too much to one very high-level abstraction.

[Yudkowsky][14:52]

On my first point, it seems to me that your claims about recursive self-improvement were off in a fairly similar way to how I think your claims about consequentialism are off - which is that they defer too much to one very high-level abstraction.

I suppose that is what it could potentially feel like from the inside to not get an abstraction. Robin Hanson kept on asking why I was trusting my abstractions so much, when he was in the process of trusting his worse abstractions instead.

[Ngo][14:51][14:53]

Explicit reflection is one possible later stage of the path; an earlier part of the path is from being optimized to do things difficult enough that you need to stop stepping on your own feet and have different parts of your thoughts work well together.

Can you explain a little more what you mean by "have different parts of your thoughts work well together"? Is this something like the capacity for metacognition; or the global workspace; or self-control; or...?

And I guess there's no good way to quantify *how* important you think the explicit reflection part of the path is, compared with other parts of the path - but any rough indication of whether it's a more or less crucial component of your view?

[Yudkowsky][14:55]

Can you explain a little more what you mean by "have different parts of your thoughts work well together"? Is this something like the capacity for metacognition; or the global workspace; or self-control; or...?

No, it's like when you don't, like, pay five apples for something on Monday, sell it for two oranges on Tuesday, and then trade an orange for an apple.

I have still not figured out the homework exercises to convey to somebody the Word of Power which is "coherence" by which they will be able to look at the water, and see "coherence" in places like a cat walking across the room without tripping over itself.

When you do lots of reasoning about arithmetic correctly, without making a misstep, that long chain of thoughts with many different pieces diverging and ultimately converging, ends up making some statement that is... still true and still about numbers! Wow! How do so many different thoughts add up to having this property? Wouldn't they wander off and end up being about tribal politics instead, like on the Internet?

And one way you could look at this, is that even though all these thoughts are taking place in a bounded mind, they are shadows of a higher unbounded structure which is the model identified by the Peano axioms; all the things being said are *true about the numbers*. Even

though somebody who was missing the point would at once object that the human contained no mechanism to evaluate each of their statements against all of the numbers, so obviously no human could ever contain a mechanism like that, so obviously you can't explain their success by saying that each of their statements was true about the same topic of the numbers, because what could possibly implement that mechanism which (in the person's narrow imagination) is The One Way to implement that structure, which humans don't have?

But though mathematical reasoning can sometimes go astray, when it works at all, it works because, in fact, even bounded creatures can sometimes manage to obey local relations that in turn add up to a global coherence where all the pieces of reasoning point in the same direction, like photons in a laser lasing, even though there's no internal mechanism that enforces the global coherence at every point.

To the extent that the outer optimizer trains you out of paying five apples on Monday for something that you trade for two oranges on Tuesday and then trading two oranges for four apples, the outer optimizer is training all the little pieces of yourself to be locally coherent in a way that can be seen as an imperfect bounded shadow of a higher unbounded structure, and then the system is powerful though imperfect *because* of how the power is present in the coherence and the overlap of the pieces, *because* of how the higher perfect structure is being imperfectly shadowed. In this case the higher structure I'm talking about is Utility, and doing homework with coherence theorems leads you to appreciate that we only know about one higher structure for this class of problems that has a dozen mathematical spotlights pointing at it saying "look here", even though people have occasionally looked for alternatives.

And when I try to say this, people are like, "Well, I looked up a theorem, and it talked about being able to identify a unique utility function from an infinite number of choices, but if we don't have an infinite number of choices, we can't identify the utility function, so what relevance does this have" and this is a kind of mistake I don't remember even coming close to making so I do not know how to make people stop doing that and maybe I can't.

[Soares][15:07]

We're already pushing our luck on time, so I nominate that we wrap up (after, perhaps, a few more Richard responses if he's got juice left.)

[Yudkowsky][15:07]

Yeah, was thinking the same.

[Soares][15:07]

As a proposed cliffhanger to feed into the next discussion, my take is that Richard's comment:

On my first point, it seems to me that your claims about recursive self-improvement were off in a fairly similar way to how I think your claims about consequentialism are off - which is that they defer too much to one very high-level abstraction.

probably contains some juicy part of the disagreement, and I'm interested in Eliezer understanding Richard's claim to the point of being able to paraphrase it to Richard's satisfaction.

[Ngo][15:08]

Wrapping up here makes sense.

I endorse the thing Nate just said.

I also get the sense that I have a much better outline now of Eliezer's views about consequentialism (if not the actual details and texture).

On a meta level, I personally tend to focus more on things like "how should we understand cognition" and not "how should we understand geopolitics and how it affects the level of pivotal action required".

If someone else were trying to prosecute this disagreement they might say much more about the latter. I'm uncertain how useful it is for me to do so, given that my comparative advantage compared with the rest of the world (and probably Eliezer's too) is the cognition part.

[Yudkowsky][15:12]

Reconvene... tomorrow? Monday of next week?

[Ngo][15:12]

Monday would work better for me.

You okay with me summarising the discussion so far to [some people — redacted for privacy reasons]?

[Yudkowsky][15:13]

Nate, take a minute to think of your own thoughts there?

[Soares:]

[Soares][15:15]

My take: I think it's fine to summarize, though generally virtuous to mark summaries as summaries (rather than asserting that your summaries are Eliezer-endorsed or w/e).

[Ngo: ]

[Yudkowsky][15:16]

I think that broadly matches my take. I'm also a bit worried about biases in the text summarizer, and about whether I managed to say anything that Rob or somebody will object to pre-publication, but we ultimately intended this to be seen and I was keeping that in mind, so, yeah, go ahead and summarize.

[Ngo][15:17]

Great, thanks

[Yudkowsky][15:17]

I admit to being curious as to what you thought was said that was important or new, but that's a question that can be left open to be answered at your leisure, earlier in your day.

[Ngo][15:17]

I admit to being curious as to what you thought was said that was important or new, but that's a question that can be left open to be answered at your leisure, earlier in your day.

You mean, what I thought was worth summarising?

[Yudkowsky][15:17]

Yeah.

[Ngo][15:18]

Hmm, no particular opinion. I wasn't going to go out of my way to do so, but since I'm chatting to [some people — redacted for privacy reasons] regularly anyway, it seemed low-cost to fill them in.

At your leisure, I'd be curious to know how well the directions of discussion are meeting your goals for what you want to convey when this is published, and whether there are topics you want to focus on more.

[Yudkowsky][15:19]

I don't know if it's going to help, but trying it currently seems better than to go on saying nothing.

[Ngo][15:20]

(personally, in addition to feeling like less of an expert on geopolitics, it also seems more sensitive for me to make claims about in public, which is another reason I haven't been digging into that area as much)

[Soares][15:21]

(personally, in addition to feeling like less of an expert on geopolitics, it also seems more sensitive for me to make claims about in public, which is another reason I haven't been digging into that area as much)

(seems reasonable! note, though, that i'd be quite happy to have sensitive sections stricken from the record, insofar as that lets us get more convergence than we otherwise would, while we're already in the area)

[Ngo: ]

(tho ofc it is less valuable to spend conversational effort in private discussions, etc.)

[Ngo: ]

[Ngo][15:22]

At your leisure, I'd be curious to know how well the directions of discussion are meeting your goals for what you want to convey when this is published, and whether there are topics you want to focus on more.

(this question aimed at you too Nate)

Also, thanks Nate for the moderation! I found your interventions well-timed and useful.

[Soares: ]

[Soares][15:23]

(this question aimed at you too Nate)

(noted, thanks, I'll probably write something up after you've had the opportunity to depart for sleep.)

On that note, I declare us adjourned, with intent to reconvene at the same time on Monday.

Thanks again, both.

[Ngo][15:23]

Thanks both 😊

Oh, actually, one quick point

Would one hour earlier suit, for Monday?

I've realised that I'll be moving to a one-hour-later time zone, and starting at 9pm is slightly suboptimal (but still possible if necessary)

[Soares][15:24]

One hour earlier would work fine for me.

[Yudkowsky][15:25]

Doesn't work as fine for me because I've been trying to avoid any food until 12:30p my time, but on that particular day I may be more caloried than usual from the previous day, and could possibly get away with it. (That whole day could also potentially fail if a minor medical procedure turns out to take more recovery than it did the last time I had it.)

[Ngo][15:26]

Hmm, is this something where you'd have more information on the day? (For the calories thing)

[Yudkowsky][15:27]

(seems reasonable! note, though, that i'd be quite happy to have sensitive sections stricken from the record, insofar as that lets us get more convergence than we otherwise would, while we're already in the area)

I'm a touch reluctant to have discussions that we intend to delete, because then the larger debate will make less sense once those sections are deleted. Let's dance around things if we can.

[Ngo: 👍] [Soares: 👍]

I mean, I can that day at 10am my time say how I am doing and whether I'm in shape for that day.

[Ngo][15:28]

great. and if at that point it seems net positive to postpone to 11am your time (at the cost of me being a bit less coherent later on) then feel free to say so at the time

on that note, I'm off

[Yudkowsky][15:29]

Good night, heroic debater!

[Soares][16:11]

At your leisure, I'd be curious to know how well the directions of discussion are meeting your goals for what you want to convey when this is published, and whether there are topics you want to focus on more.

The discussions so far are meeting my goals quite well so far! (Slightly better than my expectations, hooray.) Some quick rough notes:

- I have been enjoying EY explicating his models around consequentialism.
 - The objections Richard has been making are ones I think have been floating around for some time, and I'm quite happy to see explicit discussion on it.
 - Also, I've been appreciating the conversational virtue with which the two of you have been exploring it. (Assumption of good intent, charity, curiosity, etc.)
- I'm excited to dig into Richard's sense that EY was off about recursive self improvement, and is now off about consequentialism, in a similar way.
 - This also sees to me like a critique that's been floating around for some time, and I'm looking forward to getting more clarity on it.
- I'm a bit torn between driving towards clarity on the latter point, and shoring up some of the progress on the former point.
 - One artifact I'd really enjoy having is some sort of "before and after" take, from Richard, contrasting his model of EY's views before, to his model now.
 - I also have a vague sense that there are some points Eliezer was trying to make, that didn't quite feel like they were driven home; and dually, some pushback by Richard that didn't feel quite frontally answered.
 - One thing I may do over the next few days is make a list of those places, and see if I can do any distilling on my own. (No promises, though.)
 - If that goes well, I might enjoy some side-channel back-and-forth with Richard about it, eg during some more convenient-for-Richard hour (or, eg, as a thing to do on Monday if EY's not in commission at 10a pacific.)

[Ngo][5:40] (next day, Sep. 9)

The discussions so far are [...]

What do you mean by "latter point" and "former point"? (In your 6th bullet point)

[Soares][7:09] (next day, Sep. 9)

What do you mean by "latter point" and "former point"? (In your 6th bullet point)

former = shoring up the consequentialism stuff, latter = digging into your critique re: recursive self improvement etc. (The nesting of the bullets was supposed to help make that clear, but didn't come out well in this format, oops.)

4. Follow-ups

4.1. Richard Ngo's summary

[Ngo] (Sep. 10 Google Doc)

2nd discussion

(Mostly summaries not quotations; also ~~hasn't yet been evaluated by Eliezer~~)

Eliezer, summarized by Richard: "The A core concept which people have trouble grasping is consequentialism. People try to reason about *how* AIs will solve problems, and ways in which they might or might not be dangerous. But they don't realise that the ability to solve a wide range of difficult problems implies that an agent must be doing a powerful search over possible solutions, which is ~~the~~ a core skill required to take actions which greatly affect the world. Making this type of AI safe is like trying to build an AI that drives red cars very well, but can't drive blue cars - there's no way you get this by default, because the skills involved are so similar. And because the search process ~~is so general~~ is by default so general, ~~it'll be very hard to~~ I don't currently see how to constrain it into any particular region."

[Yudkowsky][10:48] (Sep. 10 comment)

The

A concept, which some people have had trouble grasping. There seems to be an endless list. I didn't have to spend much time contemplating consequentialism to derive the consequences. I didn't spend a lot of time talking about it until people started arguing.

[Yudkowsky][10:50] (Sep. 10 comment)

the

a

[Yudkowsky][10:52] (Sep. 10 comment)

[the search process] is [so general]

"is by default". The reason I keep emphasizing that things are only true by default is that the work of surviving may look like doing hard nondefault things. I don't take fatalistic "will happen" stances, I assess difficulties of getting nondefault results.

[Yudkowsky][10:52] (Sep. 10 comment)

it'll be very hard to

"I don't currently see how to"

[Ngo] (Sep. 10 Google Doc)

Eliezer, summarized by Richard (continued): "In biological organisms, evolution is ~~one source~~ the ultimate source of consequentialism. A ~~second~~ secondary outcome of evolution is reinforcement learning. For an animal like a cat, upon catching a mouse (or failing to do so) many parts of its brain get slightly updated, in a loop that makes it more likely to catch the mouse next time. (Note, however, that this process isn't powerful enough to make the cat a pure consequentialist - rather, it has many individual traits that, when we view them from this lens, point in the same direction.) ~~A third thing that makes humans in particular consequentialist is planning,~~ Another outcome of evolution, which helps make humans in particular more consequentialist, is planning - especially when we're aware of concepts like utility functions."

[Yudkowsky][10:53] (Sep. 10 comment)

one

the ultimate

[Yudkowsky][10:53] (Sep. 10 comment)

second

secondary outcome of evolution

[Yudkowsky][10:55] (Sep. 10 comment)

especially when we're aware of concepts like utility functions

Very slight effect on human effectiveness in almost all cases because humans have very poor reflectivity.

[Ngo] (Sep. 10 Google Doc)

Richard, summarized by Richard: "Consider an AI that, given a hypothetical scenario, tells us what the best plan to achieve a certain goal in that scenario is. Of course it needs to do consequentialist reasoning to figure out how to achieve the goal. But that's different from an AI which chooses what to say as a means of achieving its goals. I'd argue that the former is doing consequentialist reasoning without itself being a consequentialist, while the latter is actually a consequentialist. Or more succinctly: consequentialism = problem-solving skills + using those skills to choose actions which achieve goals."

Eliezer, summarized by Richard: "The former AI might be slightly safer than the latter if you could build it, but I think people are likely to dramatically overestimate how big the effect is. The difference could just be one line of code: if we give the former AI our current scenario as its input, then it becomes the latter. For purposes of understanding alignment difficulty, you want to be thinking on the level of abstraction where you see that in some sense it is the search itself that is dangerous when it's a strong enough search, rather than the danger seeming to come from details of the planning process. One particularly helpful thought experiment is to think of advanced AI as an '[outcome pump](#)' which selects from futures in which a certain outcome occurred, and takes whatever action leads to them."

[Yudkowsky][10:59] (Sep. 10 comment)

particularly helpful

"attempted explanatory". I don't think most readers got it.

I'm a little puzzled by how often you write my viewpoint as thinking that whatever I happened to say a sentence about is the Key Thing. It seems to rhyme with a deeper failure of many EAs to pass the MIRI [ITI](#).

To be a bit blunt and impolite in hopes that long-languishing social processes ever get anywhere, two obvious uncharitable explanations for why some folks may systematically misconstrue MIRI/Eliezer as believing much more than in reality that various concepts an argument wanders over are Big Ideas to us, when some conversation forces us to go to that place:

(A) It paints a comfortably unflattering picture of MIRI-the-Other as weirdly obsessed with these concepts that seem not so persuasive, or more generally paints the Other as a bunch of weirdos who stumbled across some concept like "consequentialism" and got obsessed with it. In general, to depict the Other as thinking a great deal of some idea (or explanatory thought experiment) is to tie and stake their status to the listener's view of how much status that idea deserves. So if you say that the Other thinks a great deal of some idea that isn't obviously high-status, that lowers the Other's status, which can be a comfortable thing to do.

(cont.)

(B) It paints a more comfortably self-flattering picture of a continuing or persistent disagreement, as a disagreement with somebody who thinks that some random concept is much higher-status than it really is, in which case there isn't more to done or understood except to duly politely let the other person try to persuade you the concept deserves its high status. As opposed to, "huh, maybe there is a noncentral point that the other person sees themselves as being stopped on and forced to explain to me", which is a much less self-flattering viewpoint on why the conversation is staying within a place. And correspondingly more of a viewpoint that somebody else is likely to have of us, because it is a comfortable view to them, than a viewpoint that it is comfortable to us to imagine them having.

Taking the viewpoint that somebody else is getting hung up on a relatively noncentral point can also be a flattering self-portrait to somebody who believes that, of course. It doesn't mean they're right. But it does mean that you should be aware of how the Other's story, told from the Other's viewpoint, is much more liable to be something that the Other finds sensible and perhaps comfortable, even if it implies an unflattering (and untrue-seeming and perhaps untrue) view of yourself, than something that makes the Other seem weird and silly and which it is easy and congruent for you yourself to imagine the Other thinking.

[Ngo][11:18] (Sep. 12 comment)

I'm a little puzzled by how often you write my viewpoint as thinking that whatever I happened to say a sentence about is the Key Thing.

In this case, I emphasised the outcome pump thought experiment because you said that the time-travelling scenario was a key moment for your understanding of optimisation, and the outcome pump seemed to be similar enough and easier to convey in the summary, since you'd already written about it.

I'm also emphasising consequentialism because it seemed like the core idea which kept coming up in our first debate, under the heading of "deep problem-solving patterns". Although I take your earlier point that you tend to emphasise things that your interlocutor is more skeptical about, not necessarily the things which are most central to your view. But if consequentialism isn't in fact a very central concept for you, I'd be interested to hear what role it plays.

[Ngo] (Sep. 10 Google Doc)

Richard, summarized by Richard: "There's a component of 'finding a plan which achieves a certain outcome' which involves actually solving the object-level problem of how someone who is given the plan can achieve the outcome. And there's another component which is figuring out how to manipulate that person into doing what you want. To me it seems like Eliezer's argument is that there's no training regime which leads an AI to spend 99% of its time thinking about the former, and 1% thinking about the latter."

[Yudkowsky][11:20] (Sep. 10 comment)

no training regime

...that the training regimes we come up with first, in the 3 months or 2 years we have before somebody else destroys the world, will not have this property.

I don't have any particularly complicated or amazingly insightful theories of why I keep getting depicted as a fatalist; but my world is full of counterfactual functions, not constants. And I am always aware that if we had access to a real Textbook from the Future explaining all of the methods that are actually robust in real life - the equivalent of telling us in advance about all the ReLUs that in real life were only invented and understood a few decades after sigmoids - we could go right ahead and build a superintelligence that thinks $2 + 2 = 5$.

All of my assumptions about "I don't see how to do X" are always labeled as ignorance on my part and a default because we won't have enough time to actually figure out how to do X. I am constantly maintaining awareness of this because being **wrong** about it being difficult is a major place where **hope** potentially comes from, if there's some idea like ReLUs that robustly vanquishes the difficulty, which I just didn't think of. Which does not, alas, mean that I am wrong about any particular thing, nor that the infinite source of optimistic ideas that is the wider field of "AI alignment" is going to produce a good idea from the same process that generates all the previous naive optimism through not seeing where the original difficulty comes from or what other difficulties surround obvious naive attempts to solve it.

[Ngo] (Sep. 10 Google Doc)

Richard, summarized by Richard (continued): "While this may be true in the limit of increasing intelligence, the most relevant systems are the earliest ones that are above human level. But humans deviate from the consequentialist abstraction you're talking about in all sorts of ways - for example, being raised in different cultures can make people much more or less consequentialist. So it seems plausible that early AGIs can be superhuman while also deviating strongly from this abstraction - not necessarily in the same ways as humans, but in ways that we push them towards during training."

Eliezer, summarized by Richard: "Even at the Einstein or von Neumann level these types of deviations start to subside. And the sort of pivotal acts which might realistically work require skills *significantly* above human level. I think even 1% of the cognition of an AI that can assemble advanced nanotech, thinking about how to kill humans, would doom us. Your other suggestions for pivotal acts (surveillance to restrict AGI proliferation; persuading world leaders to restrict AI development) are not politically feasible in real life, to the level required to prevent the world from ending; or else require alignment in the very dangerous domain of superhuman manipulation."

Richard, summarized by Richard: "I think we probably also have significant disagreements about geopolitics which affect which acts we expect to be pivotal, but it seems like our comparative advantage is in discussing cognition, so let's focus on that. We can build systems that outperform humans at quite a few tasks by now, without them needing search abilities that are general enough to even try to take over the world. Putting aside for a moment the question of which tasks are pivotal enough to save the world, which parts of your model draw the line between human-level chess players and human-level galaxy-colonisers, and say that we'll be able to align ones that significantly outperform us on *these* tasks before they take over the world, but not on *those* tasks?"

Eliezer, summarized by Richard: "One aspect there is domain generality which in turn is achieved through novel domain learning. One can imagine asking the question: is there a superintelligent AGI that can quickly build nanotech the way that a beaver solves building dams, in virtue of having a bunch of specialized learning abilities without it ever having a cross-domain general learning ability? But there are many, many, many things that humans do which no other animal does, which you might think would contribute a lot to that animal's fitness if there were animalistic ways to do it - e.g. mining and smelting iron. (Although comparisons to animals are not generally reliable arguments about what AIs can do - e.g. chess is much easier for chips than neurons.) So my answer is 'Perhaps, but not by default, there's a bunch of subproblems, I don't actually know how to do it right now, it's not the easiest way to get an AGI that can build nanotech.' Can I explain how I know that? I'm really not sure I can."

[Yudkowsky][11:26] (Sep. 10 comment)

Can I explain how I know that? I'm really not sure I can.

In original text, this sentence was followed by a long attempt to explain anyways; if deleting that, which is plausibly the correct choice, this lead-in sentence should also be deleted, as otherwise it paints a false picture of how much I would try to explain anyways.

[Ngo][11:15] (Sep. 12 comment)

Makes sense; deleted.

[Ngo] (Sep. 10 Google Doc)

Richard, summarized by Richard: "Challenges which are trivial from a human-engineering perspective can be very challenging from an evolutionary perspective (e.g. spinning wheels). So the evolution of animals-with-a-little-bit-of-help-from-humans might end up in very different places from the evolution of animals-just-by-themselves. And analogously, the ability of humans to fill in the gaps to help less general AIs achieve more might be quite significant."

"On nanotech: what are the most relevant axes of difference between solving protein folding and designing nanotech that, say, self-assembles into a computer?"

Eliezer, summarized by Richard: "This question seemed potentially cruxy to me. I.e., if building protein factories that built nanofactories that built nanomachines that met a certain deep and lofty engineering goal, didn't involve cognitive challenges different in kind from protein folding, we could maybe just safely go do that using AlphaFold 3, which would be just as safe as AlphaFold 2. I don't think we can do that. But it is among the more plausible advance-specified miracles we could get. At this point our last hope is that in fact the future is often quite surprising."

Richard, summarized by Richard: "It seems to me that you're making the same mistake here as you did with regards to recursive self-improvement in the AI foam debate - namely, putting too much trust in one big abstraction."

Eliezer, summarized by Richard: "I suppose that is what it could potentially feel like from the inside to not get an abstraction. Robin Hanson kept on asking why I was trusting my abstractions so much, when he was in the process of trusting his worse abstractions instead."

4.2. Nate Soares' summary

[Soares] (Sep. 12 Google Doc)

Consequentialism

Ok, here's a handful of notes. I apologize for not getting them out until midday Sunday. My main intent here is to do some shoring up of the ground we've covered. I'm hoping for skims and maybe some light comment back-and-forth as seems appropriate (perhaps similar to Richard's summary), but don't think we should derail the main thread over it. If time is tight, I would not be offended for these notes to get little-to-no interaction.

--

My sense is that there's a few points Eliezer was trying to transmit about consequentialism, that I'm not convinced have been received. I'm going to take a whack at it. I may well be wrong, both about whether Eliezer is in fact attempting to transmit these, and about whether Richard received them; I'm interested in both protests from Eliezer and paraphrases from Richard.

[Soares] (Sep. 12 Google Doc)

1. "The consequentialism is in the plan, not the cognition".

I think Richard and Eliezer are coming at the concept "consequentialism" from very different angles, as evidenced eg by Richard saying (Nate's crappy paraphrase:) "where do you think the consequentialism is in a cat?" and Eliezer responding (Nate's crappy paraphrase:) "the cause of the apparent consequentialism of the cat's behavior is distributed between its brain and its evolutionary history".

In particular, I think there's an argument here that goes something like:

- Observe that, from our perspective, saving the world seems quite tricky, and seems likely to involve long sequences of clever actions that force the course of history into a narrow band (eg, because if we saw short sequences of dumb actions, we could just get started).
- Suppose we were presented with a plan that allegedly describes a long sequence of clever actions that would, if executed, force the course of history into some narrow band.
 - For concreteness, suppose it is a plan that allegedly funnels history into the band where we have wealth and acclaim.
- One plausible happenstance is that the plan is not in fact clever, and would not in fact have a forcing effect on history.
 - For example, perhaps the plan describes founding and managing some silicon valley startup, that would not work in practice.
- Conditional on the plan having the history-funnelling property, there's a sense in which it's scary regardless of its source.
 - For instance, perhaps the plan describes founding and managing some silicon valley startup, and will succeed virtually every time it's executed, by dint of having very generic descriptions of things like how to identify and respond

- to competition, including descriptions of methods for superhumanly-good analyses of how to psychoanalyze the competition and put pressure on their weakpoints.
- In particular, note that one need not believe the plan was generated by some "agent-like" cognitive system that, in a self-contained way, made use of reasoning we'd characterize as "possessing objectives" and "pursuing them in the real world".
 - More specifically, the scariness is a property of the plan itself. For instance, the fact that this plan accrues wealth and acclaim to the executor, in a wide variety of situations, regardless of what obstacles arise, implies that the plan contains course-correcting mechanisms that keep the plan on-target.
 - In other words, plans that *manage to actually funnel history* are (the argument goes) liable to have a wide variety of course-correction mechanisms that keep the plan oriented towards *some* target. And while this course-correcting property tends to be a property of history-funneling plans, the *choice of target* is of course free, hence the worry.

(Of course, in practice we perhaps shouldn't be visualizing a single Plan handed to us from an AI or a time machine or whatever, but should instead imagine a system that is reacting to contingencies and replanning in realtime. At the least, this task is easier, as one can adjust only for the contingencies that are beginning to arise, rather than needing to predict them all in advance and/or describe general contingency-handling mechanisms. But, and feel free to take a moment to predict my response before reading the next sentence, "run this AI that replans autonomously on-the-fly" and "run this AI+human loop that replans+reevaluates on the fly", are still in this sense "plans", that still likely have the property of Eliezer!consequentialism, insofar as they work.)

[Soares] (Sep. 12 Google Doc)

There's a part of this argument I have not yet driven home. Factoring it out into a separate bullet:

2. "If a plan is good enough to work, it's pretty consequentialist in practice".

In attempts to collect and distill a handful of scattered arguments of Eliezer's:

If you ask GPT-3 to generate you a plan for saving the world, it will not manage to generate one that is very detailed. And if you tortured a big language model into giving you a detailed plan for saving the world, the resulting plan would not work. In particular, it would be full of errors like insensitivity to circumstance, suggesting impossible actions, and suggesting actions that run entirely at cross-purposes to one another.

A plan that is sensitive to circumstance, and that describes actions that synergize rather than conflict -- like, in Eliezer's analogy, photons in a

laser -- is much better able to funnel history into a narrow band.

But, on Eliezer's view as I understand it, this "the plan is not constantly tripping over its own toes" property, goes hand-in-hand with what he calls "consequentialism". As a particularly stark and formal instance of the connection, observe that one way a plan can trip over its own toes is if it says "then trade 5 oranges for 2 apples, then trade 2 apples for 4 oranges". This is clearly an instance of the plan failing to "lase" -- of some orange-needing part of the plan working at cross-purposes to some apple-needing part of the plan, or something like that. And this is also a case where it's easy to see how if a plan *is* "lasing" with respect to apples and oranges, then it is behaving as if governed by some coherent preference.

And the point as I understand it isn't "all toe-tripping looks superficially like an inconsistent preference", but rather "insofar as a plan *does* manage to chain a bunch of synergistic actions together, it manages to do so precisely insofar as it is Eliezer!consequentialist".

cf the analogy to [information theory](#), where if you're staring at a maze and you're trying to build an accurate representation of that maze in your own head, you will succeed precisely insofar as your process is Bayesian / information-theoretic. And, like, this is supposed to feel like a fairly tautological claim: you (almost certainly) can't get the image of a maze in your head to match the maze in the world by visualizing a maze at random, you have to add visualized-walls using some process that's correlated with the presence of actual walls. Your maze-visualizing process will work precisely insofar as you have access to & correctly make use of, observations that correlate with the presence of actual walls. You might also visualize extra walls in locations where it's politically expedient to believe that there's a wall, and you might also avoid visualizing walls in a bunch of distant regions of the maze because it's dark and you haven't got all day, but the resulting visualization in your head is accurate precisely *insofar* as you're managing to act kinda like a Bayesian.

Similarly (the analogy goes), a plan works-in-concert and avoids-stepping-on-its-own-toes precisely insofar as it is consequentialist. These are two sides of the same coin, two ways of seeing the same thing.

And, I'm not so much attempting to *argue* the point here, as to make sure that the *shape of the argument* (as I understand it) has been understood by Richard. In particular, the *shape of the argument* I see Eliezer as making is that "clumsy" plans don't work, and "laser-like plans" work insofar as they are managing to act kinda like a consequentialist.

Rephrasing again: we have a wide variety of mathematical theorems all spotlighting, from different angles, the fact that a plan lacking in clumsiness, is possessing of coherence.

("And", my model of Eliezer is quick to note, "this ofc does not mean that all sufficiently intelligent minds must generate very-coherent plans. If you really knew what you were doing, you could design a mind that emits plans that always "trip over themselves" along one particular axis, just as

with sufficient mastery you could build a mind that believes $2+2=5$ (for some reasonable cashing-out of that claim). But you don't get this for free -- and there's a sort of "attractor" here, when building cognitive systems, where just as generic training will tend to cause it to have true beliefs, so will generic training tend to cause its plans to lase.")

(And ofc much of the worry is that all the mathematical theorems that suggest "this plan manages to work precisely insofar as it's lasing in some direction", say nothing about which direction it must lase. Hence, if you show me a plan clever enough to force history into some narrow band, I can be fairly confident it's doing a bunch of lasing, but not at all confident which direction it's lasing in.)

[Soares] (Sep. 12 Google Doc)

One of my guesses is that Richard does in fact understand this argument (though I personally would benefit from a paraphrase, to test this hypothesis!), and perhaps even buys it, but that Richard gets off the train at a following step, namely that we *need* plans that "lase", because ones that don't aren't strong enough to save us. (Where in particular, I suspect most of the disagreement is in how far one can get with plans that are more like language-model outputs and less like lasers, rather than in the question of which pivotal acts would put an end to the acute risk period)

But setting that aside for a moment, I want to use the above terminology to restate another point I saw Eliezer as attempting to make: one big trouble with alignment, in the case where we need our plans to be like lasers, is that on the one hand we need our plans to be like lasers, but on the other hand we want them to *fail* to be like lasers along certain specific dimensions.

For instance, the plan presumably needs to involve all sorts of mechanisms for refocusing the laser in the case where the environment contains fog, and redirecting the laser in the case where the environment contains mirrors (...the analogy is getting a bit strained here, sorry, bear with me), so that it can in fact hit a narrow and distant target. Refocusing and redirecting to stay on target are part and parcel to plans that can hit narrow distant targets.

But the humans shutting the AI down is like scattering the laser, and the humans tweaking the AI so that it plans in a different direction is like them tossing up mirrors that redirect the laser; and we want the plan to fail to correct for those interferences.

As such, on the Eliezer view as I understand it, we can see ourselves as asking for a very unnatural sort of object: a path-through-the-future that is robust enough to funnel history into a narrow band in a very wide array of circumstances, but somehow insensitive to specific breeds of human-initiated attempts to switch which narrow band it's pointed towards.

Ok. I meandered into trying to re-articulate the point over and over until I had a version distilled enough for my own satisfaction (which is much like arguing the point), apologies for the repetition.

I don't think debating the claim is the right move at the moment (though I'm happy to hear rejoinders!). Things I would like, though, are: Eliezer saying whether the above is on-track from his perspective (and if not, then poking a few holes); and Richard attempting to paraphrase the above, such that I believe the arguments themselves have been communicated (saying nothing about whether Richard also buys them).

[Soares] (Sep. 12 Google Doc)

My Richard-model's stance on the above points is something like "This all seems kinda plausible, but where Eliezer reads it as arguing that we had better figure out how to handle lasers, I read it as an argument that we'd better save the world without needing to resort to lasers. Perhaps if I thought the world could not be saved except by lasers, I would share many of your concerns, but I do not believe that, and in particular it looks to me like much of the recent progress in the field of AI -- from AlphaGo to GPT to AlphaFold -- is evidence in favor of the proposition that we'll be able to save the world without lasers."

And I recall actual-Eliezer saying the following (more-or-less in response, iiuc, though readers note that I might be misunderstanding and this might be out-of-context):

Definitely, "turns out it's easier than you thought to use gradient descent's memorization of zillions of shallow patterns that overlap and recombine into larger cognitive structures, to add up to a consequentialist nanoengineer that only does nanosystems and never does sufficiently general learning to apprehend the big picture containing humans, while still understanding the goal for that pivotal act you wanted to do" is among the more plausible advance-specified miracles we could get.

On my view, and I think on Eliezer's, the "zillions of shallow patterns"-style AI that we see today, is not going to be sufficient to save the world (nor destroy it). There's a bunch of reasons that GPT and AlphaZero aren't destroying the world yet, and one of them is this "shallowness" property. And, yes, maybe we'll be wrong! I myself have been surprised by how far the shallow pattern memorization has gone (and, for instance, was surprised by GPT), and acknowledge that perhaps I will continue to be surprised. But I continue to predict that the shallow stuff won't be enough.

I have the sense that lots of folk in the community are, one way or another, saying "Why not consider the problems of aligning systems that memorize zillions of shallow patterns?". And my answer is, "I still don't expect those sorts of machines to either kill or save us, I'm still expecting that there's a phase shift that won't happen until AI systems start to be able to make plans that are sufficiently deep and laserlike to do scary stuff, and I'm still expecting that the real alignment challenges are in that regime."

And this seems to me close to the heart of the disagreement: some people (like me!) have an intuition that it's quite unlikely that figuring out how to get sufficient work out of shallow-memorizers is enough to save us, and I suspect others (perhaps even Richard!) have the sense that the aforementioned "phase shift" is the unlikely scenario, and that I'm focusing on a weird and unlucky corner of the space. (I'm curious whether you endorse this, Richard, or some nearby correction of it.)

In particular, Richard, I am curious whether you endorse something like the following:

- I'm focusing ~all my efforts on the shallow-memorizers case, because I think shallow-memorizer-alignment will by and large be sufficient, and even if it is not then I expect it's a good way to prepare ourselves for whatever we'll turn out to need in practice. In particular I don't put much stock in the idea that there's a predictable phase-change that forces us to deal with laser-like planners, nor that predictable problems in that domain give large present reason to worry.

(I suspect not, at least not in precisely this form, and I'm eager for corrections.)

I suspect something in this vicinity constitutes a crux of the disagreement, and I would be thrilled if we could get it distilled down to something as concise as the above. And, for the record, I personally endorse the following counter to the above:

- I am focusing ~none of my efforts on shallow-memorizer-alignment, as I expect it to be far from sufficient, as I do not expect a singularity until we have more laser-like systems, and I think that the laserlike-planning regime has a host of predictable alignment difficulties that Earth does not seem at all prepared to face (unlike, it seems to me, the shallow-memorizer alignment difficulties), and as such I have large and present worries.

[Soares] (Sep. 12 Google Doc)

Ok, and now a few less substantial points:

There's a point Richard made here:

Oh, interesting. Actually one more question then: to what extent do you think that explicitly reasoning about utility functions and laws of rationality is what makes consequentialists have the properties you've been talking about?

that I suspect constituted a miscommunication, especially given that the following sentence appeared in Richard's summary:

A third thing that makes humans in particular consequentialist is planning, especially when we're aware of concepts like utility

functions.

In particular, I suspect Richard's model of Eliezer's model places (or placed, before Richard read Eliezer's comments on Richard's summary) some particular emphasis on systems reflecting and thinking about their own strategies, as a method by which the consequentialism and/or effectiveness gets in. I suspect this is a misunderstanding, and am happy to say more on my model upon request, but am hopeful that the points I made a few pages above have cleared this up.

Finally, I observe that there are a few places where Eliezer keeps beeping when Richard attempts to summarize him, and I suspect it would be useful to do the dorky thing of Richard very explicitly naming Eliezer's beeps as he understands them, for purposes of getting common knowledge of understanding. For instance, things I think it might be useful for Richard to say verbatim (assuming he believes them, which I suspect, and subject to Eliezer-corrections, b/c maybe I'm saying things that induce separate beeps):

1. Eliezer doesn't believe it's impossible to build AIs that have most any given property, including most any given safety property, including most any desired "non-consequentialist" or "deferential" property you might desire. Rather, Eliezer believes that many desirable safety properties don't happen by default, and require mastery of minds that likely takes a worrying amount of time to acquire.
2. The points about consequentialism are not particularly central in Eliezer's view; they seem to him more like obvious background facts; the reason conversation has lingered here in the EA-sphere is that this is a point that many folk in the local community disagree on.

For the record, I think it might also be worth Eliezer acknowledging that Richard probably understands point (1), and that glossing "you don't get it for free by default and we aren't on course to have the time to get it" as "you can't" is quite reasonable when summarizing. (And it might be worth Richard counter-acknowledging that the distinction is actually quite important once you buy the surrounding arguments, as it constitutes the difference between describing the current playing field and laying down to die.) I don't think any of these are high-priority, but they might be useful if easy :-)

Finally, stating the obvious-to-me, none of this is intended as criticism of either party, and all discussing parties have exhibited significant virtue-according-to-Nate throughout this process.

[Yudkowsky][21:27] (Sep. 12)

From Nate's notes:

For instance, the plan presumably needs to involve all sorts of mechanisms for refocusing the laser in the case where the environment contains fog, and redirecting the laser in the case where the environment contains mirrors (...the analogy is getting a bit strained here, sorry, bear with me), so that it can in fact hit a narrow and distant target. Refocusing and redirecting to stay on target are part and parcel to plans that can hit narrow distant targets.

But the humans shutting the AI down is like scattering the laser, and the humans tweaking the AI so that it plans in a different direction is like them tossing up mirrors that redirect the laser; and we want the plan to fail to correct for those interferences.

--> GOOD ANALOGY.

...or at least it sure conveys to *me* why corrigibility is anticonvergent / anticoherent / actually *moderately strongly contrary to* and not just *an orthogonal property* of a powerful-plan generator.

But then, I already know why that's true and how it generalized up to resisting our various attempts to solve small pieces of more important aspects of it - it's not just true by weak default, it's true by a stronger default where a roomful of people at a workshop spend several days trying to come up with increasingly complicated ways to describe a system that will let you shut it down (but not steer you through time *into* shutting it down), and all of those suggested ways get shot down. (And yes, people outside MIRI now and then publish papers saying they totally just solved this problem, but all of those "solutions" are things we considered and dismissed as trivially failing to scale to powerful agents - they didn't understand what we considered to be the first-order problems in the first place - rather than these being evidence that MIRI just didn't have smart-enough people at the workshop.)

[Yudkowsky][18:56] (Nov. 5 follow-up comment)

Eg, "Well, we took a system that only learned from reinforcement on situations it had previously been in, and couldn't use imagination to plan for things it had never seen, and then we found that if we didn't update it on shut-down situations it wasn't reinforced to avoid shutdowns!"

Study Guide

This post is for students who hope to eventually work on technical [problems we don't understand](#), especially agency and AI alignment, and want to know what to study or practice.

Guiding Principles

Current alignment researchers have wildly different recommendations on paths into the field, usually correlated with the wildly different paths these researchers have themselves taken into the field. This also correlates with different kinds of work on alignment. This guide largely reflects my own path, and I think it is useful if you want to do the sort of research I do. That means fairly theoretical work (for now), very technical, drawing on models and math from a lot of different areas to understand real-world agents.

[Specializing in Problems We Don't Understand](#) lays out a general framework which guides many of the recommendations here. I'll also briefly go over some guiding principles more specific to choosing what (and how much) to study:

- Breadth over depth
- Practice generalizing concepts
- Be able to model anything
- High volume of knowledge

Breadth Over Depth

In general, study in any particular topic has decreasing marginal returns. The first exposure or two gives you the basic frames, tells you what kinds of questions to ask and what kinds of tools are available, etc. You may not remember everything, but you can at least remember what things to look up later if you need them - which is a pretty huge improvement over not even knowing that X is a thing you can look up at all!

Another way to frame this: problems-we-don't-understand rely heavily on bringing in frames and tools from other fields. (If the frames and tools of this field were already sufficient, it wouldn't be a problem-we-don't-understand in the first place.) So, you want to have a very large library of frames and tools to apply. On the other hand, you don't necessarily need very much depth in each frame or tool - just enough to recognize problems where it might apply and maybe try it out in a quick-and-dirty way.

Practice Generalizing Concepts

Bringing in frames and tools from other fields requires the ability to recognize and adapt those frames and tools for problems very different from the field in which we first learned them. So, practice generalizing concepts from one area to another is particularly important.

Unfortunately, this is not a focus in most courses. There are exceptions - applied math classes often involve applying tools in a wide variety of ways, and low-level physics courses often provide very good practice in applying a few mathematical tools to a wide variety of problems. Ultimately, though, this is something you should probably practice on your own a lot more than it's practiced in class.

Keeping a list of 10-20 hard problems in the back of your mind, and trying out each new frame or tool on one of those problems, is a particularly useful technique to practice generalization.

Be Able To Model Anything

One common pitfall is to be drawn into areas which advertise extreme generality, but are rarely useful in practice. (A lot of high-level math is like this.) On the other hand, we still want a lot of breadth, including things which are not obviously useful to whatever problem we're most interested in (e.g. alignment). After all, if the obviously-relevant tools sufficed, then it wouldn't be a problem-we-don't-understand in the first place.

To that end, it's useful to look for frames/tools which are at least useful for *something*. Keeping a list of 10-20 hard problems in the back of your mind is one useful test for this. Another useful heuristic is "be able to model anything": if there's some system or phenomenon which you're not sure how to model, even in principle, and field X has good tools for modelling it, then study field X.

This heuristic is useful for another reason, too: our intuitions for a problem of interest often come from other systems, and you never know what system will seem like a useful analogue. If we can model anything, then we always know how to formalize a model based on any particular analogy - we're rarely left confused about how to even set it up.

High Volume of Knowledge

Lastly, one place where I differ from the recommendations which I expect most current alignment researchers to give: I recommend studying a *lot*. This is based on my own experience - I've covered an awful lot of ground, and when I trace the sources of my key thoughts on alignment and agency, they come from an awful lot of places.

To that end: don't just take whatever courses are readily available. I recommend heavy use of online course material from other schools, as well as textbooks. Sometimes the best sources are a lot better than the typical source - I try to highlight any particularly great sources I know of in this post. Also, I've found it useful to "pre-game" the material even for my normal college courses - i.e. find a book or set of lectures covering similar material, and go through them before the semester starts, so that the in-person class is a second exposure rather than a first exposure. (This also makes the course a lot easier, and makes it easier overall to maintain ok grades without having to sink overly-pointless levels of effort into the class.)

Other useful tips to squeeze out every last drop:

- Skipping pre-reqs is often a good idea.

- Audit courses. This doesn't just have to be at your school - I've audited half a dozen courses at schools where I had no formal affiliation. Just walk in on the first day of class and sit down, it's usually totally fine, professors love it (since you're actually interested).

All that said, obviously **this advice is for the sort of person who is not already struggling to keep up with a more normal course load**. This advice is definitely not for everyone.

Coursework/Textbooks

With guiding principles out of the way, on to the main event: things to study. We'll start with technical foundations, i.e. the sort of stuff which might be "common core classes" at a high-end STEM college/university. Then, we'll cover topics which might be in an (imaginary) "alignment and agent foundations" degree. Finally, I'll go through a few more topics which aren't obviously relevant to alignment or agency, but are generally-useful for modelling a wide variety of real-world systems.

If I know of a particularly good source I'll link to it, but sometimes the only sources I've used are mediocre or offline. Sorry. Also, I went to Harvey Mudd College, so any references to classes there are things I did in-person.

Technical Foundations

High-School Basics

- Programming
- Calculus
- Prob/stat
- Chemistry
- Physics

If your high-school doesn't have a programming class, use a MOOC, preferably in Python. There are lots of good sources available nowadays; the "intro to programming" market is very saturated. Heck, the "intro" market is pretty saturated in all of these.

Physics and calculus go together; calculus will likely feel unmotivated without physics, and physics will have a giant calculus-shaped hole in it without calculus.

Programming

You should probably take more than one undergrad-level intro programming course, ideally using different languages. Different courses focus on very different things: low-level computer system concepts, high-level algorithms, programming language concepts, etc. Also, different languages serve very different use-cases and induce different thinking-patterns, so it's definitely worth knowing a few, ideally very different languages.

Besides basic programming fluency, you should learn:

- Basics of big-O analysis
- A conceptual understanding of how a computer works (but probably not all the low-level details)

Personally, I've used [Harvard's CS50](#), a set of intro lectures from UNSW, CS5 & CS60 at Harvey Mudd, plus a Java textbook in high school. At bare minimum, you should probably work with C/C++, Python, and a LISP variant. ([Harvard's CS50](#) is good for C/C++, [MIT has an intro in LISP](#) which is widely considered very good, and lots of courses use Python.)

It's also worthwhile to learn the basics of javascript and build a simple dynamic website at some point, but I rarely see an actual *class* in that.

Data Structures

Once you've had one or two intro programming classes, there's usually a course in data structures. It will cover things like arrays, linked lists, hash tables, trees, heaps, queues, etc. This is the bread-and-butter of most day-to-day programming.

Although the coursework may not emphasize it, I recommend building a habit of keeping a [Fermi estimate](#) of program runtime in the back of your head. I'd even say that the *main* point of learning about all these data structures is to make that Fermi estimate.

Linear Algebra

Linear algebra is the main foundational tool we need for mathematically modelling anything with a lot of dimensions, i.e. [our world](#). In practice, most of the matrices we use are either:

- First or second derivatives of high-dimensional functions, or
- Data on which we calculate correlations/run linear regressions.

Alas, when first studying the subject, it will probably be very abstract and you won't see good examples of what it's actually used for. (It is useful, though - I last used linear algebra yesterday, when formulating an [abstraction](#) problem as an eigenproblem.)

Linear algebra took me many passes to learn well. I read three textbooks and took two in-person courses (from different schools) in linear algebra, then took another two courses (also from different schools) in linear systems. Out of all that, the only resource I strongly recommend is [Boyd's lectures on linear dynamical systems](#), probably after one or two courses in linear algebra. I also hear [Linear Algebra Done Right](#) is good as an intro, but haven't used it personally. [MIT's lectures](#) are probably very good, though sadly I don't think they were online back when I was learning the subject.

If you take more advanced math/engineering, you'll continue to learn more linear algebra, especially in areas like linear control theory, Fourier methods, and PDEs.

Mechanics & Differential Equations

Mechanics (usually a physics class) and differential equations (a math class) are the two courses where you go from mostly-not-knowing-how-to-model-most-things to mostly-having-some-idea-how-to-model-most-things-at-least-in-principle. In particular, I remember differential equations as the milestone where I transitioned from feeling like there were small islands of things I knew how to model mathematically, to small islands of things I *didn't* know how to model mathematically, at least in principle. (I had taken some mechanics before that.)

I took all my mechanics in-person, but I hear the [Feynman Lectures](#) are an excellent source. For differential equations, I used [MIT's lectures](#). You will need some linear algebra for differential equations (at least enough to not run away screaming at the mention of eigenvalues), though not necessarily on the first pass (some schools break it up into a first course without linear algebra and then a second course with it).

Multivariate Calculus

In principle, multivariate calculus is what makes linear algebra useful. Unfortunately, multivariate calculus courses in my experience are a grab-bag of topics, some which are quite useful, others of which are pretty narrow.

The topics in my ideal course in multivariate calculus would be:

- Tensor notation
- Tensor & matrix calculus
- Gradients & gradient descent optimization
- Hessians & Newton's Method optimization
- Jacobians & Newton's Method root finding
- Constrained optimization & Lagrange multipliers
- Jacobian determinants & multivariate coordinate transformations for integrals
- Wedge products
- Conservative vector fields & potentials

About half of these are covered very well in Boyd's convex optimization course (see below). The rest you may have to pick up piecemeal:

- [Tensor notation](#) you can just adopt for yourself and practice; it's very useful for ML, continuum mechanics, and general relativity
- [Matrix calculus](#) you'll pick up if you need to hand-code fast gradient calculations for optimization or simulation problems
- Jacobian determinants will come up whenever a high-dimensional integral requires a coordinate change. Play around with it and then practice it when it's needed.
- Wedge products are useful whenever an integral is over a multi-dimensional surface in some higher-dimensional space; when you write "dx dy dz" in an integral, that's secretly a wedge product. Again, play around with it and then practice it when it's needed.
- Conservative vector fields you'll see a lot in electricity & magnetism (as well as specific techniques for them)

Convex Optimization

Linear algebra, as we use it today, is a relatively recent development:

The separate linear algebra course became a standard part of the college mathematics curriculum in the United States in the 1950s and 60s and some colleges and universities were still adding the course in the early 1970s. ([source](#))

Fifty years ago, linear algebra was new. What new things today will be core technical classes in another fifty years, assuming a recognizable university system still exists?

I think convex optimization is one such topic.

Boyd is the professor to learn this from, and [his lectures](#) are excellent. This is one of my strongest not-already-standard recommendations in this post.

Bayesian Probability

Another topic which is on the short list for “future STEM core”. I don’t have a 101-level intro which I can personally vouch for - [Yudkowsky’s intro](#) is popular, but you’ll probably need a full course in probability before diving into the more advanced stuff.

You can get away with a more traditional probability course and then reading Jaynes (see below), which is what I did, but a proper Bayesian probability course is preferred if you can find a good one.

Microeconomics

Economics provides the foundations for a ton of agency models.

Any standard 101-level course is probably fine. Lean towards more math if possible; for someone doing all the other courses on this list, there’s little reason not to jump into the math.

Proofs

Alignment theory involves proving things, so you definitely need to be comfortable writing proofs.

To the extent that proof-writing is taught, it’s unfortunately often taught in Analysis 1, which is mostly-useless in practice other than the proof skills. (There are lots of useful things in analysis, but mostly I recommend you skip the core “analysis” courses and learn the useful parts in other classes, like theoretical mechanics or math finance or PDEs or numerical analysis.) Pick up proof skills elsewhere if you can; you’ll have ample opportunity to practice in all the other classes on this list.

Agency and Alignment “Major”

AI & Related Topics

Intro AI

Mostly this course will provide a first exposure to stuff you'll study more later. Pay attention to relaxation-based search in particular; it's a useful unifying framework for a lot of other things.

I took [Norvig & Thrun's MOOC](#) when it first came out, which was quite good. Russell & Norvig's [textbook](#) appears to cover similar material.

Causality

Turns out we *can* deduce causality from correlation, it just requires more than two variables. More generally, causal models are the main "language" you need to speak in order to efficiently translate intuitions about the world into Bayesian probabilistic models.

Yudkowsky has a [decent intro](#), although you definitely need more depth than that. [Pearl's books](#) are canonical; [Koller & Friedman](#) are unnecessarily long but definitely cover all the key pieces. Koller has a [coursera course](#) covering similar material, which would probably be a good choice.

Jaynes

Jaynes' [Probability Theory: The Logic Of Science](#) is a book for which I know no substitute. It is a book on Bayesian probability theory by the leading Bayesian probability theorist of the twentieth century; other books on the topic look sloppy by comparison. There are insights in this book which I have yet to find in any other book or course.

At the bare minimum, read chapters 1-4 and 20. I've read it cover-to-cover, and found it immensely valuable.

Information Theory

Information theory is a powerful tool for translating a variety of intuitions into math, especially agency-adjacent intuitions.

I don't know of any really good source on information theory, but I do remember that there's one textbook from about 50 years ago which is notoriously terrible. If you find yourself wading through lots of analysis, put the book down and find a different one.

I have used a set of "[Information Theory and Entropy](#)" lectures from MIT, which are long but have great coverage of topics, especially touching on more physics-flavored stuff. I also use [Cover & Thomas](#) as a reference, mainly because it has good chapters on Kelly betting and portfolio optimization.

Godel Escher Bach

Another book for which I know no substitute. [Godel Escher Bach](#) is... hard to explain. But it's a fun read, you should read it cover-to-cover, and you will have much better conceptual foundations for thinking about self-reflection and agency afterwards.

ML

Obviously some hands-on experience with ML is useful for anyone working on AI, even theoretical work - current systems are an important source of "data" on agency, same as biology and economics and psychology/neuroscience. Also, it's one of those classes which brings together a huge variety of technical skills, so you can practice all that linear algebra and calculus and programming.

Unfortunately, these days there's a flood of ML intros which don't have any depth and just tell you how to call magic black-boxes. For theoretical agency/alignment work, that's basically useless; understanding what goes on inside of these systems is where most of the value comes from. So look for a course/book which involves building as much as possible from scratch.

You might also consider an "old-school" ML course, from back before deep learning took off. I used [Andrew Ng's old lectures](#) back in the day. A lot of the specific algorithms are outdated now, but there's a lot of math done automatically now which we used to have to do by hand (e.g. backpropagating gradients). Understanding all that math is important for theory work, so doing it the old-fashioned way a few times can be useful.

Other than understanding the internals of deep learning algorithms, I'd also recommend looking into the new generation of probabilistic programming languages (e.g. [Pyro](#)), and how they work.

Algorithms

I've heard a saying that you can become a great programmer either by programming for ten years, or by programming for five years and taking an algorithms class. For theory work, a solid understanding of algorithms is even more important - we need to know what's easy, what's hard, and be able to recognize easy vs hard things in the wild.

Algorithms courses vary a lot in what they cover, but some key things which you definitely want:

- Dynamic programming. I've used [one of Bellman's books](#) on the subject, which was excellent.
- NP-completeness & reductions. You need to be able to recognize the kinds-of-problems which are usually NP-complete, and be able to prove that they're NP-complete if necessary.
- Relaxation-based search (i.e. A* search), if you haven't already covered in depth in an intro AI course

Depending on how much depth you want on the more theoretical parts, [Avi Wigderson has a book](#) with ridiculously deep and up-to-date coverage, though the writing is often overly abstract.

Numerical Algorithms

Numerical algorithms are the sort of thing you use for simulating physical systems or for numerical optimization in ML. Besides the obvious object-level usefulness, many key ideas of numerical algorithms (like sparse matrix methods or condition numbers) are really more-general principles of world modelling, which for some reason people don't talk about much until you're up to your elbows in actual numerical code.

Courses under names like "numerical algorithms", "numerical analysis", or "scientific computing" cover various pieces of the relevant material; it's kind of a grab-bag.

Biology

For purposes of agency and alignment work, biology is one of the main sources of evolved agenty systems. It's a major source of intuitions and qualitative data for my work (and hopefully quantitative data, some day). Also, if you want to specialize in problems-we-don't-understand more generally, biology will likely be pretty central.

The two most important books to read are Alon's [Design Principles of Biological Circuits](#), and the [Bionumbers book](#). The former is about the surprising extent to which evolved biological systems have unifying human-legible design principles (I have a review [here](#)). The latter is an entire book of Fermi estimates, and will give you lots of useful intuitions and visualizations for what's going on in cells.

I also strongly recommend a course in synthetic biology. I used a set of lectures which I think were a pilot for [this course](#).

Economics

Like biology, economics is a major source of intuitions and data on agenty systems. Unlike biology, it's also a major source of mathematical models for agenty systems. I think it is very likely that a successful theory of the foundations of agency will [involve](#) market-like structures and math.

I don't know of any very good source on the "core" market models of modern economics beyond the 101 level. I suspect that [Stokey, Lucas and Prescott](#) does a good job (based on other work by the authors), but I haven't read it myself. I believe you'd typically find this stuff in a first-year grad-school microeconomics course.

If you want to do this the hard way: first take convex optimization (see above), then try to solve the N Economists Problem.

N economists walk into a bar, each with a utility function and a basket of goods. Compute the equilibrium distribution of goods.

This requires making some reasonably-general standard economic assumptions (concave increasing utility functions, rational agents, common knowledge, Law of One Price).

Learning it the hard way takes a while.

Once you have the tools to solve the N Economists problem (whether from a book/course or by figuring it out the hard way), the next step along the path is “[dynamic stochastic general equilibrium](#)” models and “[recursive macro](#)”. (These links are to two books I happen to have, but there are others and I don’t have any reason to think these two are unusually good.) You probably do *not* need to go that far for alignment work, but if you want to specialize in problems-we-don’t-understand more generally, then these tools are the cutting-edge baseline for modelling markets (especially financial markets).

Game Theory

Game theory is the part of economics most directly relevant to alignment and agency, and largely independent of market models, so it gets its own section.

You might want to take an intro-level course if you don’t already know the basics (e.g. what a Nash equilibrium is), but you might just pick that up somewhere along the way. Once you know the very basics, I recommend two books. First, [Games and Information](#) by Eric Rasmusen. It’s all about games in which the players have different information - things like principal-agent problems, signalling, mechanism design, bargaining, etc. This is exactly the right set of topics to study, which largely makes up for a writing style which I don’t particularly love. (You might be able to find a course which covers similar material.)

The other book is Thomas Schelling’s [Strategy of Conflict](#), the book which [cousin_it summarized as](#):

Forget rationalist Judo: this is rationalist eye-gouging, rationalist gang warfare, rationalist nuclear deterrence. Techniques that let you win, but you don’t want to look in the mirror afterward.

For this book, I don’t know of any good substitute.

Control Theory

Control systems are all over the place in engineered devices. Even your thermostat needs to not be too sensitive in blasting out hot/cold air in response to cold/hot temperatures, lest we get amplifying hot/cold cycles. It’s a simple model, but even complex AI systems (or biological systems, or economic systems) can be modeled as control systems.

You’ll probably pick up the basics of linear control theory in other courses on this list (especially linear dynamical systems). If you want more than that, [one of Bellman’s books](#) on dynamic programming and control theory is a good choice, and [these lectures on underactuated control](#) are really cool. This is another category where you only need the very basics for thinking about alignment and agency, but more advanced knowledge is often useful for a wide variety of problems.

Dynamical Systems

Chaos is conceptually fundamental to all sorts of “complex systems”. It’s quite central to [my own work on abstraction](#), and I wouldn’t be at all surprised if it has other

important applications in the theory of agency.

There's many different classes where you might pick up an understanding of chaos, but a course called "Nonlinear Dynamical Systems" (or something similar) is the most likely bet.

Statistical Mechanics

Probably my biggest mistake in terms of undergraduate coursework was not taking statistical mechanics. It's an alternative viewpoint for all the probability theory and information theory stuff, and it's a viewpoint very concretely applied in everyday situations. Some of it is physics-specific, but it's an ongoing source of key ideas nonetheless.

If you can learn Bayesian stat mech, that's ideal, although it's not taught that way everywhere and I don't know of a good textbook. (If you want a pretty advanced and dense book, [Walter T Grandy](#) is your guy, but that one is a bit over my head.)

The Sequences

In case nobody mentioned it yet, you probably want to read the [sequences](#), including [these two](#). They're long, but they cover a huge amount of important conceptual material, and they're much lighter reading than technical textbooks.

Useful In General, But Not So Much For Alignment

This section is intended for people who want to specialize in technical problems-we-don't-understand more generally, beyond alignment. It contains courses which I've found useful for a fairly broad array of interesting problems, but less so for alignment specifically. I won't go into as much depth on these, just a quick bullet list with one-sentence blurbs and links.

- Theoretical Mechanics. Using Newton's laws for everything gets messy in more complicated systems; this course covers cleaner methods. [Susskind's lectures](#) are good.
- Quantum. If you have an itching desire to know how it works, I strongly recommend [The Quantum Challenge](#) as a starting point. That book covers the conceptually-"weird" parts much better than most courses.
- Electromagnetism. This is the more theory-heavy part of E&M, circuits is more practical. [Griffiths](#) is the standard textbook, and is quite good.
- Electronic circuits. I used [MIT's 6.002 lectures](#), which were fun.
- Digital logic/VLSI/etc. This is the class where you design a simple computer CPU starting from transistors and wires.
- Systems programming. The gnarly parts of programming - dealing with the OS and low-level code, databases, networks, etc.
- Parallel/asynchronous programming. Self explanatory.
- SQL. Also self explanatory.

- Graphics (esp. Procedural Graphics). Games and animation are one of the places where people need really robust, fast, realistic simulations of all sorts of things, which makes it a really cool area to practice lots of technical skills.
- Robotics. Another fun area to practice lots of technical skills.
- Modular arithmetic, polynomial rings, and related algorithms (polynomial multipoint, GCD, Chinese remainder). Powerful tools for certain kinds of algorithmic problems; might be scattered across a few different classes.
- Materials 101. [MIT has some really fun lectures](#).
- Continuum mechanics (i.e. Elastics & Fluid Mechanics). Core tools for modelling solids and fluids, respectively.
- Math Finance. Ito calculus in particular is a very useful tool. [Hull](#) is the standard text; any course using that text will likely cover similar material
- Fourier. Generally a useful tool for linear PDEs, and the backbone of fast convolutions (as in “convolutional neural network”). Somewhat old-school at this point.
- PDEs. Nonlinear PDEs and Numerical PDEs are usually separate classes, and are also quite useful (the former for qualitative understanding of nonlinear-specific phenomena like [shocks](#), the latter for simulation).
- Complex analysis. These tools sure do seem powerful, but I haven’t gotten much use out of them in practice. Not sure if that’s just me or not.

Final Thoughts

That was a lot. It took me roughly eight hours of typing just to write it all out, and a lot longer than that to study it all.

With that in mind: **you absolutely do not need to study all of this**. It’s a sum, not a logical-and. The more you cover, the wider the range of ideas you’ll have to draw from. It’s not like everything will magically click when you study the last piece; it’s just a long gradual accumulation.

If there’s one thing which I don’t think this list conveys enough, it’s the importance of actually playing around with all the frames and tools and trying them out on problems of your own. See how they carry over to new applications; see how to use them. Most of the things on this list I studied because they were relevant to one problem or another I was interested in, and I practiced by trying them out on those problems. Follow things which seem interesting, things for which you already have applications in mind, and you’ll learn them better. More advanced projects will practice large chunks of this list all at once. In large part, the blurbs here were meant to help suggest possible applications and stoke your interest.

Oh, one more thing: practice writing clear explanations and distillations of technical ideas. It’s a pretty huge part of alignment and agency research in practice. I hear [blog posts explaining the technical stuff you’re learning](#) are pretty good for that - and also a good way to visibly demonstrate your own understanding.

Attempted Gears Analysis of AGI Intervention Discussion With Eliezer

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

Recently, a discussion of potential AGI interventions and potential futures [was posted to LessWrong](#). The picture Eliezer presented was broadly consistent with my existing model of Eliezer's model of reality, and most of it was also consistent with my own model of reality.

Those two models overlap a lot, but they are different and my model of Eliezer strongly yells at anyone who thinks they shouldn't be different that Eliezer wrote a technically not infinite but rather very large number of words explaining that you need to think for real and evaluate such things for yourself. On that, our models definitely agree.

It seemed like a useful exercise to reread the transcript of Eliezer's discussion, and explicitly write out the world model it seems to represent, so that's what I'm going to do here.

Here are some components of Eliezer's model, directly extracted from the conversation, rewritten to be third person. It's mostly in conversation order but a few things got put in logical order.

Before publishing, I consulted with Rob Bensinger, who helped refine several statements to be closer to what Eliezer actually endorses. I explicitly note the changes where they involve new info, so it's clear what is coming from the conversation and what is coming from elsewhere. In other places it caused me to clean up my wording, which isn't noted. It's worth pointing out that the corrections often pointed in the 'less doom' direction, both in explicit claims and in tone/implication, so chances are this comes off as generally implying more doom than is appropriate.

1. Nate rather than Eliezer, but offered as a preface with $p \sim 0.85$: AGI is probably coming within 50 years. Rob notes that Eliezer may or may not agree with this timeline, and that it shortens if you condition on 'unaligned' and lengthens conditional on 'aligned.'
2. By default this AGI will come from something similar to some of today's ML paradigms. Think enormous inscrutable floating-point vectors.
3. AGI that isn't aligned ends the world.
4. AGI that isn't aligned carefully and on purpose isn't aligned, period.
5. It may be possible to align an AGI carefully and have it not end the world.
6. Right now we don't know how to do it at all, but in theory we might learn.
7. The default situation is an AGI system arises that can be made more powerful by adding more compute, and there's an extended period where it's not aligned yet and if you add too much compute the world ends, but it's possible that if you had enough time to work on it and no one did that, you'd have a shot.
8. More specifically, when combined with other parts of the model detailed later: "I think we're going to be staring down the gun of a completely inscrutable model that would kill us all if turned up further, with no idea how to read what goes on inside its head, and no way to train it on humanly scrutable and safe and

humanly-labelable domains in a way that seems like it would align the superintelligent version, while standing on top of a whole bunch of papers about “small problems” that never got past “small problems”.”

9. If we don't learn how to align an AGI via safety research, nothing else can save us period.
10. Thus, all scenarios where we win are based on a technical surprising positive development of unknown shape, and all plans worth having should assume such a surprising positive development is possible in technical space. In the post this is called a ‘miracle’ but this has misleading associations – it was not meant to imply a negligible probability, only surprise, so Rob suggested changing it to ‘surprising positive development.’ Which is less poetic and longer, but I see the problem.
11. Eliezer does know a lot of ways *not* to align an AGI, which is helpful (e.g. Edison knew a lot of ways not to build a light bulb) but also isn't good news.
12. Carefully aligning an AGI would at best be slow and difficult, requiring years of work, even if we did know how.
13. Before you could hope to finish carefully aligning an AGI, someone else with access to the code could use that code to end the world. Rob clarifies that good info security still matters and can meaningfully buy you time, and suggests this: “By default (absent strong op-sec and research closure), you should expect that before you can finish carefully aligning an AGI, someone else with access to the code could use that code to end the world. Likewise, by default (absent research closure and a large technical edge), you should expect that other projects will independently figure out how to build AGI shortly after you do.”
14. There are a few players who we might expect to choose *not* to end the world like Deepmind or Anthropic, but only a few. There are many actors, each of whom might or might not end the world in such a spot (e.g. home hobbyists or intelligence agencies or Facebook AI research), and it only takes one of them.
15. Keeping the code and insights involved secret and secure over an extended period is a level of social technology no ML group is close to having. I read the text as making the stronger claim that we lack the social technology for groups of sufficient size to keep this magnitude of secret for the required length of time, even with best known practices.
16. Trying to convince the folks that would otherwise destroy the world that their actions would destroy the world isn't impossible on some margins, so in theory some progress could be made, and some time could be bought, but not enough to buy enough time.
17. Most reactions to such problems by such folks, once their attention is drawn to them, would make things worse rather than better. Tread carefully or not at all, and trying to get the public into an uproar seems worse than useless.
18. Trying to convince various projects to become more closed rather than open is possible, and (as per Rob) a very good idea if you would actually succeed, but insufficient.
19. Trying to convince various projects to join together in the endgame, if we were to get to one, is possible, but also insufficient and (as per Rob) matters much less than becoming more closed now.
20. Closed and trustworthy projects are the key to potentially making technical progress in a safe and useful way. There needs to be a small group that can work on a project and that wouldn't publish the resulting research or share its findings automatically with a broader organization, via sufficiently robust subpartitions.
21. Anthropic in particular doesn't seem open to alternative research approaches and mostly wants to apply More Dakka, and doesn't seem open to sufficiently robust subpartitions, but those could both change.

22. Deepmind in particular is a promising potential partner if they could form the required sufficiently robust subpartitions, even if Demis must be in the loop.
23. OpenAI as a concept (rather than the organization with that name), is a maximally bad concept almost designed to make the playing field as unwinnable as possible, details available elsewhere. Of course, the organization itself *could* change (with or without being renamed to ClosedAI).
24. More generally, publishing findings burns the common resource ‘time until AGI’ and the more detail you publish about your findings along {quiet internal result -> announced and demonstrated result -> paper describing how to get the announced result -> code for the result -> model for the result} the more of it you burn, but the more money and prestige the researchers get for doing that.
25. One thing that would be a big win would be *actual* social and corporate support for subpartitioned projects that didn’t publish their findings, where it didn’t cost lots of social and weirdness points for the researchers, thus *allowing* researchers to avoid burning the commons.
26. Redwood Research (RR) is a new research organization that’s going to try and do alignment experiments on toy problems to learn things, in ways people like Eliezer think are useful and valuable and that they wish someone would do. Description not directly from Eliezer but in context seems safe to assume he roughly agrees.
27. Previously (see Hanson/Eliezer FOOM debate) Eliezer thought you’d need recursive self-improvement *first* to get fast capability gain, and now it looks like you can get fast capability gain without it, for meaningful levels of fast. This makes ‘hanging out’ at interesting levels of AGI capability at least *possible*, since it wouldn’t automatically keep going right away.
28. An AGI that was above humans in *all* respects would doubtless FOOM anyway, but if ahead in only some it might not.
29. Trying to set special case logic to tell AGIs to believe false generalizations with a lot of relevance to mapping or steering the world won’t work, they’d notice and fix it.
30. Manipulating humans is a convergent instrumental strategy.
31. Hiding what you are doing is a convergent instrumental strategy.
32. Eliezer expects that when people are trying to stomp out convergent instrumental strategies by training at a safe dumb level of intelligence, this will not be effective at preventing convergent instrumental strategies at smart levels of intelligence.
33. You *have to* train in safe domains because if you train in unsafe domains you die, but the solutions you find in safe domains won’t work in unsafe domains.
34. Attempts to teach corrigibility in safe regimes are unlikely to generalize well to higher levels of intelligence and unsafe regimes.
35. Explanation of above part 1: Higher levels of intelligence involve qualitatively new thought processes and things being way out of training distribution.
36. Explanation of above part 2: Corrigibility is ‘anti-natural’ in a certain sense that makes it incredibly hard to, eg, exhibit any coherent planning behavior (“consistent utility function”) which corresponds to being willing to let somebody else shut you off, without incentivizing you to actively manipulate them to shut you off).
37. Trying to hardcode nonsensical assumptions or arbitrary rules into an AGI will fail because a sufficiently advanced AGI will notice that they are damage and route around them or fix them (paraphrase).
38. You only get one shot, because the first miss kills you, and your chances of pulling many of these things off *on the first try* is basically zero, unless (Rob suggests this) you can basically ‘read off’ what the AI is thinking. Nothing like

this that involves black boxes ever works the first time. Alignment is hard largely because of ‘you only get one shot.’

39. Nothing we can do with a safe-by-default AI like GPT-3 would be powerful enough to save the world (to ‘commit a pivotal act’), although it might be *fun*. In order to use an AI to save the world it needs to be powerful enough that you need to trust its alignment, which doesn’t solve your problem.
40. Nanosystems are definitely possible, if you doubt that read Drexler’s *Nanosystems* and perhaps *Engines of Creation* and think about physics. They’re a core thing one could and should ask an AI/AGI to build for you in order to accomplish the things you want to accomplish.
41. No existing suggestion for “Scalable Oversight” seems to solve any of the hard problems involved in creating trustworthy systems.
42. An AGI would be able to argue for/‘prove’ arbitrary statements to the satisfaction of humans, including falsehoods.
43. Furthermore, an unaligned AGI powerful enough to commit pivotal acts should be assumed to be able to hack any human foolish enough to interact with it via a text channel.
44. The speedup step in “iterated amplification and distillation” will introduce places where the fast distilled outputs of slow sequences are not true to the original slow sequences, because gradient descent is not perfect and won’t be perfect and it’s not clear we’ll get any paradigm besides gradient descent for doing a step like that.
45. The safety community currently is mostly bouncing off the hard problems and are spending most of their time working on safe, easy, predictable things that guarantee they’ll be able to publish a paper at the end. Actually-useful alignment research will tend to be risky and unpredictable, since it’s advancing the frontier of our knowledge in a domain where we have very little already-accumulated knowledge.
46. Almost all other work is either fully useless, almost entirely predictable, or both.
47. Paul Christiano is trying to have real foundational ideas, and they’re all wrong, but he’s one of the few people trying to have foundational ideas at all; if we had another 10 of him, something might go right.
48. Chris Olah is going to get far too little done far too late but at least is trying to do things on a path to doing anything at all.
49. Stuart Armstrong did some good work on further formalizing the shutdown problem, an example case in point of why corrigibility is hard, which so far as I know is still resisting all attempts at solution.
50. Various people who work or worked for MIRI came up with some actually-useful notions here and there, like Jessica Taylor’s expected utility quantilization.
51. We need much, much more rapid meaningful progress than this to have any chance, and it’s not obvious how to do that, or how to use money usefully. Money by default produces more low-quality work, and low-quality work slash solving small problems rather than the hard problems isn’t *quite* useless but it’s not going to get us where we need to go.
52. The AGI approaches that matter are the ones that scale, so they probably look less like GPT-2 and more like Alpha Zero, AlphaFold 2 or in particular Mu Zero.
53. Proving theorems about the AGI doesn’t seem practical. Even if we somehow managed to get structures far more legible than giant vectors of floats, using some AI paradigm very different from the current one, it still seems like huge key pillars of the system would rely on non-fully-formal reasoning.
54. Zvi infers this from the text, rather than it being text directly, and it’s possible it’s due to conflating things together and wasn’t intended: A system that is mathematically understood and you can prove lots of stuff about it is not on the

table at this point. Agent Foundations is a failure. Everything in that direction is a failure.

55. Even if you could prove what the utility function was, getting it to actually represent a human-aligned thing when it counts still seems super hard even if it doesn't involve a giant inscrutable vector of floats, and it probably does involve that.
56. Eliezer agrees that it seems plausible that the good cognitive operations we want do not *in principle* require performing bad cognitive operations; the trouble, from his perspective, is that generalizing structures that do lots of good cognitive operations will automatically produce bad cognitive operations, especially when we dump more compute into them; "you can't bring the coffee if you're dead". No known way to pull this off.
57. Proofs mostly miss the point. Prove whatever you like about that Tensorflow problem; it will make no difference to whether the AI kills you. The properties that can be proven just aren't related to safety, no matter how many times you prove an error bound on the floating-point multiplications. It wasn't floating-point error that was going to kill you in the first place.

Now to put the core of that into simpler form, and excluding non-central details, in a more logical order.

Again, this is my model of Eliezer's model, statements are not endorsed by me, I agree with many but not all of them.

1. Claim from Nate rather than Eliezer, unclear if Eliezer agrees: AGI is probably coming ($p \sim 85\%$) within 50 years.
2. AGI that is not aligned ends the world.
3. Safely aligning a powerful AGI is difficult.
4. Humanity only gets one shot at this. If we fail, we die and can't try again.
5. Almost nothing ever succeeds on its first try.
6. We currently have no idea how to do it at all.
7. Current alignment methods all fail and we don't even have good leads to solving the hard questions that matter.
8. AIs weak enough to be safe-by-default lack sufficient power to solve these problems.
9. It would take a surprising positive technical development to find a way to do alignment at all.
10. So all reasonable plans to align an AGI assume at least one surprising positive and technical development.
11. Current pace of useful safety research is much slower than needed to keep pace with capabilities research.
12. Even if we did get a surprising positive technical development that let us find a way to proceed, it would probably take additional years to do that rather than turn the AGI on and end the world. Rob Bensinger clarifies that Eliezer's exact stance is instead: "An aligned advanced AI created by a responsible project that is hurrying where it can, but still being careful enough to maintain a success probability greater than 25%, will take the lesser of (50% longer, 2 years longer) than would an unaligned unlimited superintelligence produced by cutting all possible corners."
13. That's because the AGI we need to align is likely an enormous inscrutable pile of floating-point vectors. Which makes it harder.
14. AGI likely comes from algorithms that scale with compute, so less like GPT-X and more like Mu Zero.

15. Such algorithms have to be aligned somehow before anyone scales them up too much, since that would end the world.
16. Rob's rewording: In the meantime, it's likely that the code and/or conceptual insights would leak out, absent a large, competent effort to prevent this. No leading ML organization currently seems to be putting in the required effort.
Zvi's note: I interpreted the relevant claim here as something stronger, that humanity lacks the social technology to do more than probabilistically postpone such a leak even under best practices given the likely surrounding conditions, and that no leading organizations are even doing anything resembling or trying to resemble best practices.
17. If it were to leak, someone somewhere would run the code and end the world. There are people who *probably* would know better than to scale it up and end the world, like Deepmind and Anthropic, but it wouldn't take long for many others to get the code, and then it only takes one someone who didn't know better (like an intelligence agency) to end the world anyway.
18. Most people working on safety are working on small problems rather than hard problems, or doing work with predictable outcomes, because incentives, and are therefore mostly useless (or worse).
19. There are exceptions (Paul Christiano, Chris Olah, Stuart Armstrong and some MIRI-associated people) but they are exceptions and we need vastly more of them.
20. AI work that is shared or published accelerates AGI.
21. The more details are shared, the more acceleration happens.
22. Everyone publishes anyway, in detail, because incentives.
23. Incentive changes to fix this would need to be sufficiently robust that not publishing wouldn't hurt your career prospects or cost you points, or they won't work.
24. Fixing incentives on publishing, and otherwise making more things more closed, would be helpful.
25. Ability to do subpartititioned/siloed projects within research organizations (including Deepmind and Anthropic), that would actually stay meaningfully secret, would be helpful.
26. Research that improves interpretability a lot (like Chris Olah is trying to do, but with faster progress) would be very helpful. Creative new deep alignment ideas (similar to Paul Christiano's work in depth and novelty, but not in the Paul-paradigm) would be very helpful.
27. Certain kinds of especially-valuable alignment experiments using present-day ML systems, like the experiments run by Redwood Research, would be helpful.
28. Nanotechnology is definitely physically doable and a convergent instrumental strategy, see Drexler.
29. Manipulating humans is a convergent instrumental strategy.
30. Hiding what you are doing is a convergent instrumental strategy.
31. Higher intelligence AGIs use qualitatively new thought processes that lie outside your training distribution.
32. An unaligned AGI would be able to hack any human foolish enough to read its text messages or other outputs, 'prove' arbitrary statements to human satisfaction, etc.
33. [Corrigibility is 'anti-natural'](#) and incredibly hard. Corrigibility solutions for less intelligent AGIs won't transfer to higher intelligence AGIs.
34. 'Scalable oversight' as proposed so far doesn't solve any of the hard problems.
35. 'Iterated amplification and distillation' based on gradient descent would be imperfect and the nice properties you're trying to preserve would fail. Currently we have no alternate approach.
36. Agent Foundations and similar mathematical approaches seem to be dead ends.

37. ‘Good’ cognitive operations grouped together and scaled automatically produce ‘bad’ cognitive operations. You can’t deliver the coffee if you’re dead, etc.
38. Getting a fixed utility function into an AGI at all is super hard+, getting a utility function to represent human values is super hard+, giant floating point vectors make both harder still.
39. The stuff you can prove doesn’t prove anything that matters, the stuff that would prove anything that matters you can’t prove.
40. Special case nonsensical logic or arbitrary rules will be interpreted by an AGI as damage and routed around.
41. Recursive self-improvement seems not required for fast capability gains. This means having a powerful but not self-improving or world-ending AGI at least possible.
42. An AGI better at everything than humans FOOMs anyway.

Worth noting that the more precise #12 is *substantially more optimistic* than 12 as stated explicitly here.

Looking at these 42 claims, I notice my inside view mostly agrees, and would separate them into:

Inside view disagreement but seems plausible: 1

Inside view lacks sufficient knowledge to offer an opinion: 28 (I haven’t looked for myself)

Inside view isn’t sure: 8 (if we add ‘using current ideas’ move to strong agreement), 13, 36

Weak inside view agreement – seems probably true not counting Eliezer’s opinion, but I wouldn’t otherwise be confident: 7, 9, 10, 22, 34, 35, 40

Strong inside view agreement: 2, 3, 4, 5, 6, 11, 12 (original version would be weak agreement, revised version is strong agreement), 14 (conditional on 13), 15, 16 (including the stronger version), 17, 18, 19 (in general, not for specific people), 20, 21, 23, 24, 25, 26, 27, 29, 30, 31, 32, 33, 37, 38, 39 (unless a bunch of other claims also break first), 41, 42

Thus, I have inside view agreement (e.g. I substantively agree with this picture without taking into account anyone’s opinion) on 37 of the 42 claims, including many that I believe to be ‘non-obvious’ on first encounter.

That leaves 5 remaining claims.

For 28 (Nanotechnology) I think it’s probably true, but I notice I’m counting on others models of the technology that would be involved, so I want to be careful to avoid information cascade, but my outside view strongly agrees.

For 8 (Safe AIs lack the power to save us) would require a surprising positive development for it to be wrong, in the sense that no currently proposed methods seem like they’d work. But I notice I instinctively remain hopeful for such a development, and for a solution to be found. I’m not sure how big a disagreement exists here, there might not be one.

That leaves 1 (85% AGI by 2070), 13 (AGI is a giant pile of floating-point vectors) and 36 (Agent Foundations and similar are dead ends) which are largely the same point of

doubt.

Which is basically this: I notice my inside view, while not confident in this, continues to not expect current methods to be sufficient for AGI, and expects the final form to be more different than I understand Eliezer/MIRI to think it is going to be, and that the AGI problem (not counting alignment, where I think we largely agree on difficulty) is 'harder' than Eliezer/MIRI think it is.

For 36 (Agent Foundations) in particular: I notice a bunch of people in the comments saying Agent Foundations isn't/wasn't important, and seemed like a non-useful thing to pursue, and if anything I'm on the flip side of that and am worried it and similar things were abandoned too quickly rather than too late. It's a case of 'this probably will do nothing and look stupid but might do a lot or even be the whole ballgame' and that being hard to sustain even for a group like MIRI in such a spot, but being a great use of resources in a world where things look very bad and all solutions assume surprising (and presumably important) positive developments. Everybody go deep.

For 1 (probability of AGI) in particular: I think in addition to *probably* thinking inside view that AGI is harder than Eliezer/MIRI think it is, I also think civilization's dysfunctions are more likely to disrupt things and make it increasingly difficult to do anything at all, or anything new/difficult, and also collapse or other disasters. I know Nate Sores explicitly rejects this mattering much, but it matters inside view to me quite a bit. I don't have an inside view point estimate, but if I could somehow bet utility (betting money really, really doesn't work here, at all) and could only bet once, I notice I'd at least buy 30% and sell 80%, or something like that.

Also, I noticed two interrelated things that I figured are worth noting from the comments:

1. In the comments to the OP that Eliezer's comments about small problems versus hard problems got condensed down to 'almost everyone working on alignment is faking it.' I think that is not only uncharitable, it's importantly a wrong interpretation, and motivated by viewing the situation (unconsciously?) through the lens of a battle over status and authority and blame, rather than how to collectively win from a position on a game board. The term 'faking' here is turning a claim of 'approaches that are being taken mostly have epsilon probability of creating meaningful progress' to a social claim about the good faith of those doing said research, and then interpreted as a social attack, and then therefore as an argument from authority and a status claim, as opposed to pointing out that such moves don't win the game and we need to play to win the game. I see Eliezer as *highly sympathetic* to how this type of work ends up dominating, and sees the problem as structural incentives that need to be fixed (hence my inclusion of 'because incentives' above) combined with genuine disagreement about the state of the game board. 'Faking it' is shorthand for 'you know what you're doing isn't real/useful and are doing it anyway' and introduces the accusation that leads to the rest of the logical sequence, or something. And Eliezer kind of wrote a whole sequence about exactly this, which I consider so important that a quote from it is my Twitter bio.
2. Eliezer is being accused of making an argument from authority using authority he doesn't deserve, or in a way that is disruptive, and I assume every time he sees that or anyone saying anything like "Eliezer thinks X therefore it's irrational to not think X too" or "Who are you (or am I) to disagree with the great Eliezer?" he's tearing his hair out that he spent years writing the definitive book about why "think for yourself, shmuck" is the way to go. I feel his pain, to a lesser

degree. I had a conversation last Friday where my post on the ports where I explain how I want to generate a system of selective amplification where everyone thinks for themselves in levels so as to amplify true and useful things over untrue and useless things was interpreted (by a rather smart and careful reader!) as a request to have a norm that people amplify messages without reading them carefully or evaluating whether they seemed true, the way certain camps do for any messages with the correct tribal coloring. And again, arrrggghh, pull hair out, etc.

Concentration of Force

This essay began as part one of a longer piece. Part one is standalone and "timeless." Part two is focused on the local dynamics of the EA/rationality/longtermist communities and LessWrong in November of 2021. Following wise advice from Zack_M_Davis, I've split them into two separate posts. Nevertheless, I recommend that people intending to read both seriously consider reading them back-to-back, so that the content of this one is fresh in the mind. It's both something of a prerequisite and also relevantly context-setting.

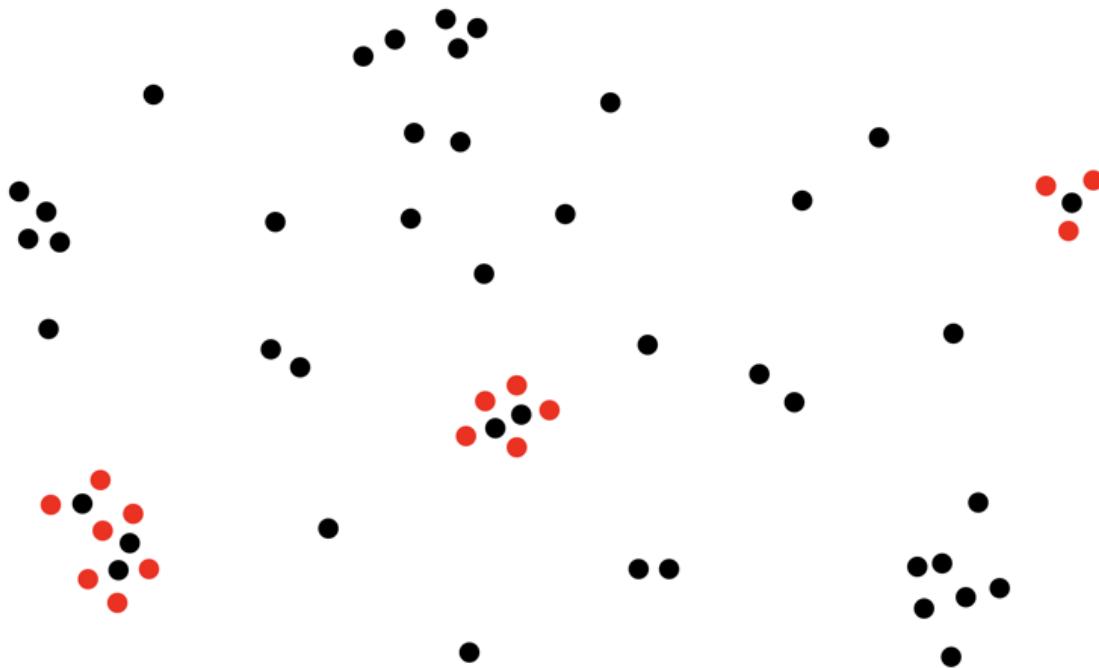
Introduction

Concentration of force is a military concept (sometimes referred to as "mass"). It used to be concentration of forces, until innovations like machine guns and cruise missiles made gathering all of your actual personnel together into more of a liability.

The idea is simple. Essentially, there is a difference between *relevant moments* and *irrelevant moments*. Battles and non-battles, moments of engagement and moments between engagements.

At each *relevant moment*, you want to project *locally superior or overwhelming force*. Perhaps this means having the most soldiers/guns/tanks/planes actually present, or perhaps this just means having the right missiles pointed in the right directions.

If you are good at coordination and maneuver, you can concentrate force in most or every engagement, and consistently win even against an overall larger or more powerful opponent. This is how guerrilla warfare works—you choose the time and place of conflict in order to ensure that you outnumber the enemy in each specific encounter, and you fade into the mists before their reinforcements arrive.



Red wins each of the depicted engagements handily.

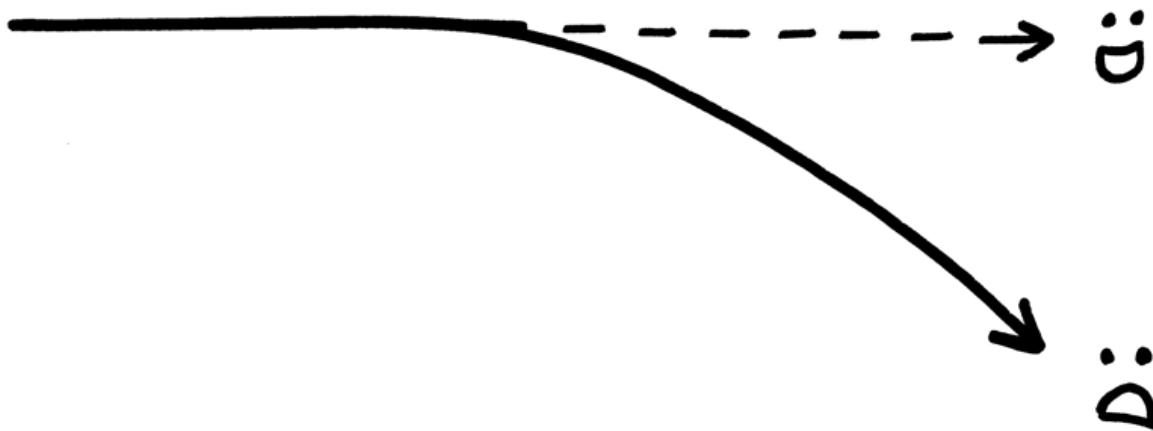
I claim that the Grey Tribe generally, and rationalists/longtermists/EAs more specifically, and LessWrong the website and community even *more* specifically, are systematically and spectacularly failing at *concentration of force*. That none of those groups puts anything like sufficient strategic energy into *ensuring that critical mass coheres at crucial moments*, and that each would benefit from optimizing their ability to do so quickly, reliably, and effectively, and from *thinking* in terms of concentration of force as a matter of habit.

I anticipate a (reasonable) objection along the lines of "mistake theory rather than conflict theory!" or "generous tit-for-tat rather than vengeful tit-for-tat!" and I assert a further subclaim that the above advice is every bit as relevant for nonviolent and nonconfrontational frames. Actual conflicts are a vanishingly small subset of the times when concentration of force is a relevant principle; the "force" in question could just as easily be e.g. "calm, generous, charitable, level-headed, clear-minded, skilled communicators arriving at exactly the moment when things were about to become wastefully contentious and adversarial."

(Indeed, that's a preview of the recommendation I have for LessWrong specifically.)

TAPs: A Motivating Example

There's a picture I tend to draw quite frequently when giving people crash courses in CFAR-esque rationality, and it looks like this:



The idea behind the picture is that you're trucking along, living a generally good and happy life, and then *something happens*, and you find yourself in the sad timeline. You ate an entire package of Oreos, despite intending to lose weight. You got in another fight with your romantic partner, despite really not wanting to. You road raged, you failed to finish the presentation before the deadline, you spent all evening on Reddit instead of thinking about your research, you somehow never called them back and now it's too awkward.

The point of showing people this picture is to draw their attention to two key facts:

1. For most goals and values, there *actually exists* a moment-of-departure from a path consistent with the positive outcome and a path consistent with the negative one. There is an identifiable point at which one of these outcomes becomes distinctly more likely than the other.

2. The paths tend to get farther and farther apart over time. It's rare that one *instantaneously* and *unrecoverably* leaps from 😊 to 😞; most of the time, there is a shift in trajectory and one's prognosis worsens as continued-progress-along-the-wrong-path compounds.

You could think of the distance between the dotted and solid lines as a measure of the *total effort required* to make it back to the better timeline. The quicker you notice that you're on the wrong track, the shorter the distance back to the right one. The less time you've spent *accelerating* in the wrong direction, the less inertia you have to overcome.

Which leads to one of the key actionable insights of [trigger-action plans](#) (known in the literature as implementation intentions): there are times where the total effort required is zero, or close enough—e.g. simply catching the moment when you *would* have made the unfortunate switch, and then not doing so. In many, many cases, an epsilon of prevention is worth an omega of cure.

One of the toy examples for CFAR's TAPs class is a person struggling with their sweet tooth:

How hard is it to stop eating Oreos once the package is open on the table in front of you?

Very. I can do it sometimes but absolutely can't count on it as a reliable strategy.

How hard is it to only take a few Oreos out of the package in the first place? Put them on a plate, and leave the package in the cupboard?

Still pretty hard.

How hard is it to not go to the cupboard at all, when you're in the grips of an Oreo craving?

Still pretty hard.

What about not buying the Oreos, when you're standing in the supermarket aisle?

Easier, but still not easy; the allure of the package is strong.

What about not going down that aisle in the first place?

Better, but still feels iffy. Still feels like it requires effort, like I'll be fighting myself the whole time I'm in the grocery store.

What about, at the moment of grabbing a shopping cart, pausing to ask whether this is an Oreos-type shopping trip, or not?

There we go.

For the (real) person in the example, triggering off of a stimulus outside the grocery store provided *sufficient distance* from the reality-distortion field of the Oreos that it was possible to make a sober yes-or-no call and *stick to it*, without ongoing effort or indecision. To sort of fortify against the urge, before it had even appeared.

This is an instance of effective *concentration of force*, and it illustrates the key point—that it (often) doesn't take much. It just takes a little *in the right place*.

Many people try to solve their TAPs-shaped problems by shotgunning effort all over the place, and most would benefit from asking themselves:

If I had only *thirty total seconds per day* of conscious awareness and available willpower, and otherwise would be stuck on autopilot and following my own personal path of least

resistance, would I be able to solve this problem?

The answer is "yes" far more often than people (who usually haven't actually tried checking) tend to think.

There is a background assumption baked into all of this that is rarely made explicit, and defended explicitly, and that is that the little stuff *actually matters*. Like keeping an extremely heavy rock balanced on its tip—it can be done with very little strength, as long as you *keep nudging* it back toward its equilibrium, never allowing it to build up momentum.

The converse of "it doesn't take much [to make things go well]" is that it doesn't take much to make them go *badly*, either. There are steep slopes and feedback loops in both directions.

Amazon Rankings: A Case Study

Readers will be able to remind me whether the cartoonist in this three-quarters-remembered anecdote is Ryan North (of Dinosaur Comics) or Zach Weinersmith (of SMBC); I was unable to dig up the details.

At some point early in the past decade, though, one of these two men was attempting to publish their very first Actual Book™, and had a clever scheme to leverage their existing (relatively small) audience. They posted a message saying approximately the following:

"Hey, everybody, I've got a book coming out soon. It'll be available on Amazon, and if you were planning to buy it, do me a favor and buy it between the hours of [time] and [time] on [specific day]."

They went on to explain that, because of the way Amazon's algorithms were structured, if enough people purchased the book during a small enough window, they could punch their way right onto the automated bestseller list, which would *then* catch the attention of the broader public (both because many, many more people would see the listing, and also because that's one heck of a story).

In this day and age, such social engineering schemes come across as somewhat passé, but at the time, this was a fairly unprecedeted hack, taking advantage of a relatively underexploited fulcrum. As I recall, it actually worked—with just a few hundred or low-thousands of purchases, the book *did* rocket to the front page, and did quite well afterward as a result.

It wasn't that our plucky cartoonist commanded a huge army of supporters. It was the fact that he *knew what to do with the small number he had*. By effectively concentrating the available force, he achieved an outsized effect, and he did so on purpose.

(For another example of this principle in practice, look [here](#).)

The Culture War is a Guerrilla War

It's not the case that there is always a *single* decisive moment. Sometimes, the relevant quality is the ability to *repeatedly* concentrate force.

Many LWers will already be familiar with the concept of [evaporative cooling](#) as applied to small subcultures (the essay is short, and worth a read if you haven't encountered it before; it's 98th-percentile in my opinion).

The metaphor of evaporative cooling also works to explain shifts in the Overton window of the larger context culture.

Consider, as a case study, the practice of parents in small towns and cities leaving their young children in the car for a few minutes while they run into the grocery store. This is a behavior my own parents engaged in, as well as the parents of approximately all of my friends and classmates circa 1990. The base rate of disaster on this activity was very low, as far as I can tell.

However, a few of those rare disasters memorably captured broad attention, and "leaving your kids in the car" acquired a slightly disreputable tinge.

As a result, those parents who were some combination of:

- Most anxious about their children's safety
- Least in-need of the benefits that leaving-your-kids-in-the-car provides, and
- Most sensitive to social disapproval

... dropped off, and stopped doing it. Some, no doubt, stopped entirely, while others just cut back on the margin.

This meant that the population of parents who *continued* to leave their kids in the car had a slightly higher proportion of parents who were:

- Slightly less attentive to their children's safety
- Slightly more invested in their ability to run into the store without their kids, and
- Slightly less responsive to social pressure

Which, in combination, had the effect of making the overall class of [kids left in cars] slightly more dangerous in actual fact, while also raising the heat in the discourse *about* the behavior (because those still actively defending it were more threatened by the prospect of its outlaw).

Thus, after a little time, the *next* layer of reasonable moderates found themselves slightly less comfortable being on Team Leave Your Kids In The Car, and stopped doing it, or at least stopped defending it in public.

Fewer respectable defenders; fewer responsible practitioners. The tinge of disrepute strengthened, with reason. The base rate of the behavior dropped further in response. The people *still* engaging in it were yet more desperate, die-hard, and bull-headed, which peeled *another* layer of moderates away. Another incident or two occurred, confirming the suspicion that the behavior itself was fundamentally dangerous, and that the people engaging in it were generically irresponsible. The situation polarized. The middle ground dissolved.

Eventually, all that remained were two tiny, armed camps made up of the small number of people still invested in shouting about it, flying the flags PROTECT CHILDREN and SAVE OUR FREEDOMS.

And for everyone who was *just trying to go about their daily lives*, it was no longer worth it to leave their kids in the car, even when it was eminently safe and reasonable to do so. It wasn't worth the risk, it wasn't worth the hassle, it wasn't worth the reputational damage and the dirty looks and the off chance of someone calling Child Protective Services and *really* ruining your day. Running errands *simply got harder* for the median and modal parents, but it was cheaper to pay the cost of keeping your kids with you or hiring a babysitter than to put forth the *extraordinary* amount of effort it would take to reclaim that tiny patch of territory in the name of sanity and reasonableness (not least because any such campaign would *first* have to do a ton of work just to differentiate itself from the crazies and their counterproductive enthusiasm).

The above story is a little too pat, and abstracts away some important detail, but it *does* gesture adequately in the direction of a very real phenomenon. When Something Goes Wrong, what usually happens is *not* that our society sits down and says "ah, here's a situation where we don't actually have clear, legible, defensible norms and policy. Let's assess the tradeoffs and come to a sensible consensus."

Instead, norms evolve in response to incentives that are often *locally* overwhelming at every point. When a Concerned Citizen™ spots a pair of four-year-olds in the car in the parking lot of the local grocery store and starts shouting about it and calls the police, each other *individual customer* has more to lose by getting involved at all than by simply turning the other way.

Meanwhile, in any random group of a thousand Americans, there will be far more who are moved by the immediate and viscerally salient stimulus of kids-who-look-like-they-might-need-protecting than those moved by the more distant and abstract harm done to norms of non-panic and non-interference. And among those who *might* be predisposed to object, there will be many who will flinch away in recognition that the Concerned Citizen could likely effectively paint them as Someone Who Doesn't Care About Kids (or at least an apologist for such), which is a substantially more powerful social weapon than Someone Who Doesn't Care About The Long-Term Ramifications Of Small Failures To Put Things In Perspective And The Tendency Of Those Small Failures To Compound.

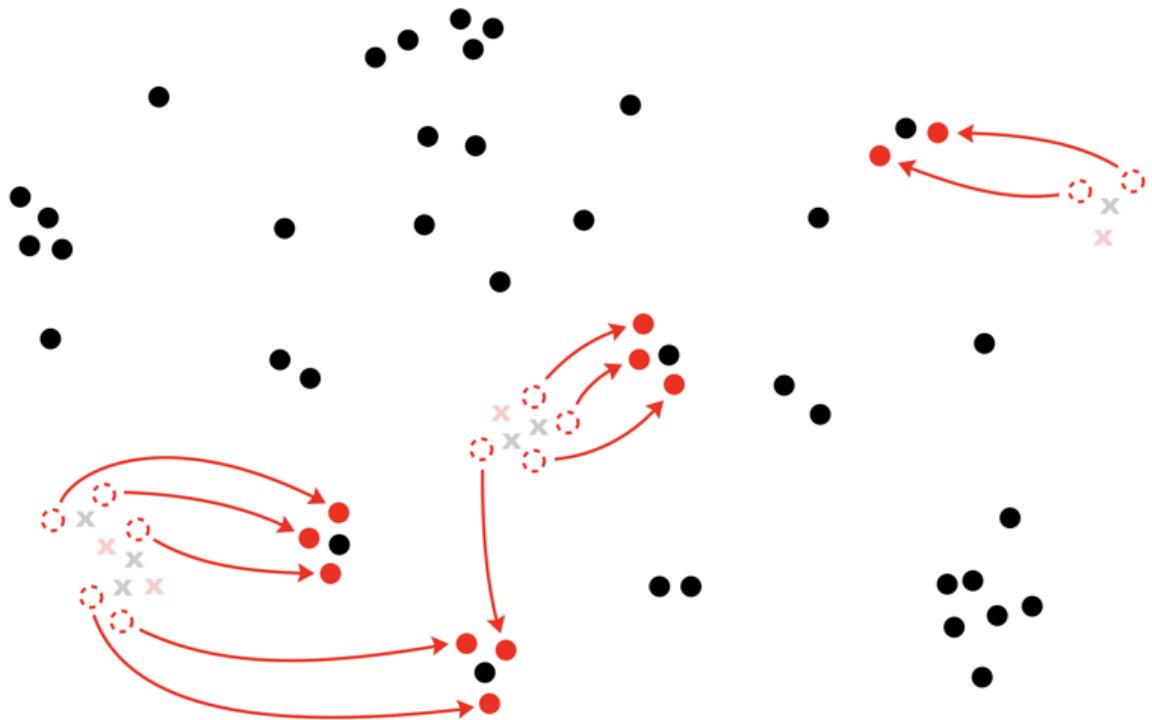
As for the police officer, or any other authority figure who arrives on the scene—many of them will know in their heart of hearts that the situation presents no real threat or concern, but it's one thing to *know that in your heart of hearts*, and it's another thing to *dismiss the concern and disperse the crowd on your own authority*. Doing so just shifts the crosshairs onto your own head—you're now the one on the hook if the Concerned Citizen is committed to making a stink, or if one of your superiors disagrees with your assessment, or if—heaven forbid—something bad actually *does* happen, after you stuck your neck out and gave your own personal stamp of approval—hmmm—*on second thought, let's just have you talk to CPS, they're the ones who are actually qualified to handle these sorts of things*—

In short, the process *driving* the evolution of the norm is a guerrilla war. The forces of reason and restraint and quantitative analysis are everywhere outnumbered in practice; it matters little that [a supermajority of people would agree that most kids-left-in-the-car scenarios pose essentially zero risk] if you cannot count on them to actually show up in the parking lot when you're under fire. Team Histrionics can more effectively *concentrate force*, and so Team Histrionics wins far more battles than it loses.

(Other examples of this dynamic include the evaporation of casual physical affection between heterosexual male friends, the disappearance of a wide range of adult mentorship relationships with children and teenagers, and the general rise of fear-of-litigiousness and the resulting increase in self-protective bureaucratic red tape.)

Morale and Momentum

Of course, in a literal military conflict, even victories of concentrated force come with a cost. Troops die or are injured, and are not easily replaced. Ammunition is spent. Equipment breaks down or is destroyed. In the original diagram at the top of this essay, the balance of power was 43(B) : 15(R). If red were to successfully take out the six surrounded black units, at the cost of (say) four of its own, the new balance would be 37(B) : 11(R). Iterate a few more times, and even if red continues to be half again or more as effective as black, they'll still nevertheless eventually be unable to muster decisive force in any given engagement.



Surviving red forces maneuvering for a second round of engagements.



The resulting balance of force after the second round: 33(B) : 9(R).

In the world of social warfare, though—especially in the digital age—this doesn't really

happen.

In social warfare, quantities like total available personnel or total materiel are largely replaced by quantities like *morale* or *zealotry*.

And morale and zealotry *anti-attract*. They snowball, rather than depleting. Success breeds enthusiasm. Each local victory of red over black energizes and inspires, leading to an increased rather than decreased willingness-to-engage in the future. Anonymity and easy access makes "showing up" extremely low-cost and often quite high-reward.

Symmetrically, defeat sets up a discouragement spiral. For every hold-the-line true believer shouting about the potential power of the silent majority, there are three or five or ten others looking at the situation and noticing that *hey, even though most people agree with this point, it nevertheless seems to be a bad idea to raise this flag?*

(And that fact itself is pretty demoralizing.)

It does not take many instances of either calling for help and not really getting it, or seeing someone else do so, to set up a self-fulfilling narrative that drastically reduces the rate of people on the black team even bothering to try.

Cancel Culture, Abridged

The Cancel Culture Essay™ will be released on some other day, but it's worth noting that cancellations as a class are just *straightforwardly* an instance of concentration of force (and their relative effectiveness a pretty strong endorsement for the principle). A highly motivated minority coordinates to ensure that anyone who runs afoul of the shared goal will receive a mountain of headache far in excess of anything they're accustomed to or capable of dealing with, and most people ~~buckle under the pressure~~

Actually, scratch that; most people can see *what's coming* and simply choose to *get out of the way*—like the filibuster, the threat is sufficiently credible that it usually doesn't actually have to be carried out.

This is true for both right-wing and left-wing cancellations; agnostic to the justification or righteousness of any given cancellation on either side of the culture war, it seems difficult to claim with a straight face that *most* of the orgs and groups withdrawing their support from various individuals are doing so because the leaders of those orgs and groups *personally* care.

Some absolutely do. That much is clear.

But the reasonable prior is that *most* of them simply do not want the headache. They do not want the negative press, they do not want the protestors, they do not want to have to explain to their shareholders why they're being dragged on Twitter when they could have just nipped this in the bud, *why would you choose this hill to die on, from what I heard it sounds like the guy is kind of a dirtbag anyway*—

The trouble with fighting for human freedom is that one spends most of one's time defending scoundrels. For it is against scoundrels that oppressive laws are first aimed, and oppression must be stopped at the beginning if it is to be stopped at all.

Many people quite reasonably do not care to publicly spend their own resources defending scoundrels, or people who unfortunately *resemble* scoundrels. Far easier to have *one* awkward conversation, where you say "Hey, look, I'm really sorry, I get that this sucks for you, but like—you get it, right? It's not that I'm unsympathetic, it's just that—"

(Gestures vaguely)

There have been relatively few cancellations where, if votes could have been cast anonymously across the entire population, a supermajority or even a *straight* majority would have been strongly in favor of so-and-so losing whatever position or status they held.

But once it becomes clear that there is a mobilized group willing not only to punish, but also to *punish non-punishers*—and once it's clear that that group can in fact swiftly and effectively concentrate force—it's no surprise that most of the non-punishers go dark. It doesn't matter that they outnumber the zealots ten to one—[it's a stag hunt, and absent reliable coordination, the only reasonable choice is rabbit.](#)

Principle, In Brief

The general lesson, I hope, is clear. To restate it:

It's often not *how much* force you can bring to bear, so much as whether you can apply that force *effectively*.

The effectiveness of force application often depends on its concentration—on whether you can amass locally superior force at the actual decisive moment.

(Both in cases where there is a single decisive moment, as in the Amazon example, or many such moments, as when cultural norms are in flux.)

Attention to this principle is generally lacking, and individuals and groups seeking to be more effective would do well to take the following advice:

0. Take seriously the idea that very small shifts in momentum really can actually snowball; do not assume *by default* that noise will swamp small influences and do not be dismissive of small interventions *just because they are small*.
1. Look for moments when small applications of force will be unusually effective; *prioritize* interventions according to how amenable they are to even quite small expenditures of resources.
2. Do what you can to make whatever-constitutes-force in the domain of your choosing *mobile* and *responsive*, such that it can be concentrated very quickly. The relevant moment is not always predictable in advance, and the "side" which can more reliably cohere decisive force faster will win more battles (many of them before they even start).

For the locally relevant essay that was originally part two, [click here](#).

Yudkowsky and Christiano discuss "Takeoff Speeds"

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

This is a transcription of Eliezer Yudkowsky responding to Paul Christiano's [Takeoff Speeds](#) live on Sep. 14, followed by a conversation between Eliezer and Paul. This discussion took place after Eliezer's [conversation](#) with Richard Ngo.

Color key:

Chat by Paul and Eliezer Other chat Inline comments

5.5. Comments on "Takeoff Speeds"

[Yudkowsky][10:14] (Nov. 22 follow-up comment)

(This was in response to an earlier request by Richard Ngo that I respond to Paul on Takeoff Speeds.)

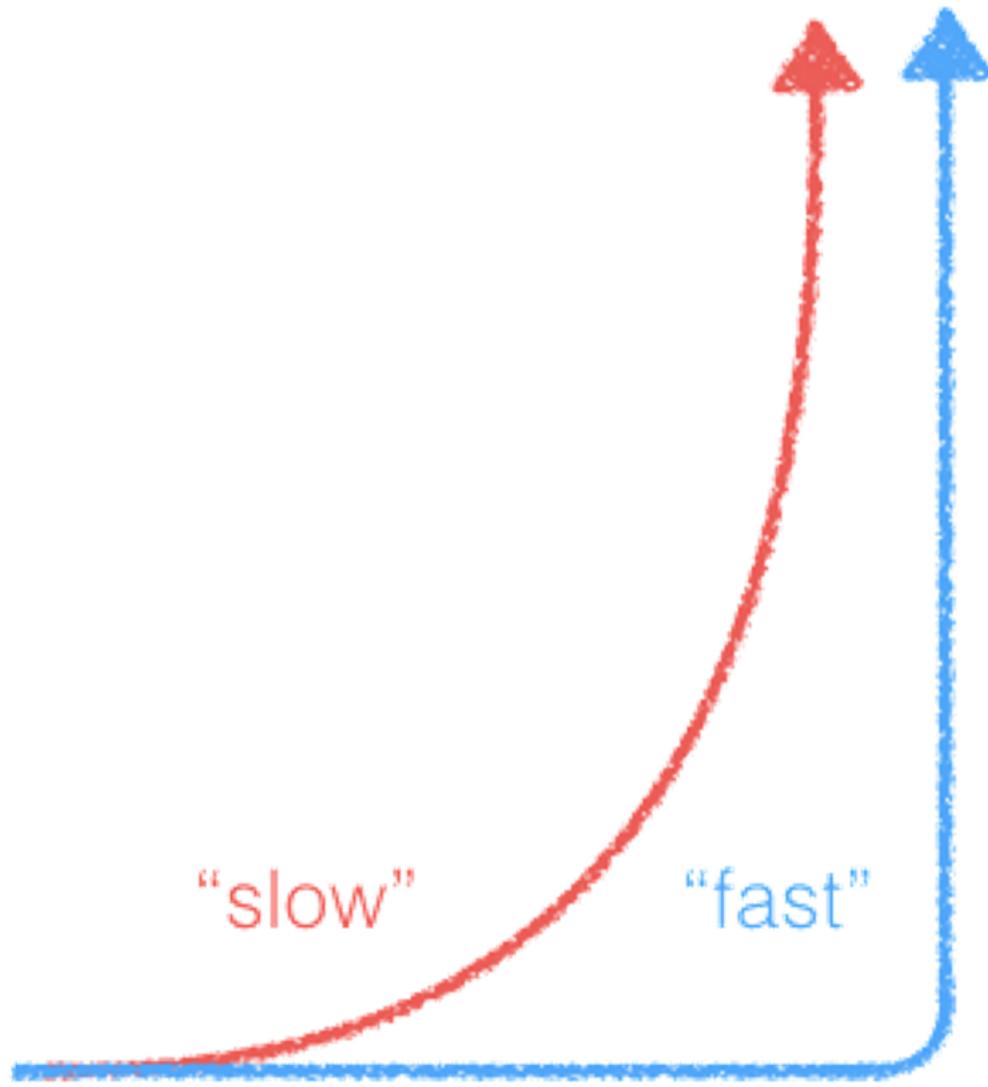
[Yudkowsky][16:52]

maybe I'll try liveblogging some <https://sideways-view.com/2018/02/24/takeoff-speeds/> here in the meanwhile

Slower takeoff means faster progress

[Yudkowsky][16:57]

The main disagreement is not about what will happen once we have a superintelligent AI, it's about what will happen *before* we have a superintelligent AI. So slow takeoff seems to mean that AI has a larger impact on the world, sooner.



It seems to me to be disingenuous to phrase it this way, given that slow-takeoff views usually imply that AI has a large impact later relative to right now (2021), even if they imply that AI impacts the world "earlier" relative to "when superintelligence becomes reachable".

"When superintelligence becomes reachable" is *not* a fixed point in time that doesn't depend on what you believe about cognitive scaling. The correct graph is, in fact, the one where the "slow" line starts a bit before "fast" peaks and ramps up slowly, reaching a high point later than "fast". It's a nice try at reconciliation with the imagined Other, but it fails and falls flat.

This may seem like a minor point, but points like this do add up.

In the fast takeoff scenario, weaker AI systems may have significant impacts but they are nothing compared to the "real" AGI. Whoever builds AGI has a decisive strategic advantage. Growth accelerates from 3%/year to 3000%/year without stopping at 30%/year. And so on.

This again shows failure to engage with the Other's real viewpoint. My mainline view is that growth stays at 5%/year and then everybody falls over dead in 3 seconds and the world gets transformed into paperclips; there's never a point with 3000%/year.

Operationalizing slow takeoff

[Yudkowsky][17:01]

There will be a complete 4 year interval in which world output doubles, before the first 1 year interval in which world output doubles.

If we allow that consuming and transforming the solar system over the course of a few days is "the first 1 year interval in which world output doubles", then I'm happy to argue that there won't be a 4-year interval with world economic output doubling before then. This, indeed, seems like a massively overdetermined point to me. That said, again, the phrasing is not conducive to conveying the Other's real point of view.

I believe that before we have incredibly powerful AI, we will have AI which is merely very powerful.

Statements like these are very often "true, but not the way the person visualized them". Before anybody built the first critical nuclear pile in a squash court at the University of Chicago, was there a pile that was almost but not quite critical? Yes, one hour earlier. Did people already build nuclear systems and experiment with them? Yes, but they didn't have much in the way of net power output. Did the Wright Brothers build prototypes before the Flyer? Yes, but they weren't prototypes that flew but 80% slower.

I guarantee you that, whatever the *fast* takeoff scenario, there will be some way to look over the development history, and nod wisely and say, "Ah, yes, see, this was not unprecedented, here are these earlier systems which presaged the final system!" Maybe you could even look back to today and say that about GPT-3, yup, totally presaging stuff all over the place, great. But it isn't transforming society because it's not over the social-transformation threshold.

AlphaFold presaged AlphaFold 2 but AlphaFold 2 is good enough to start replacing other ways of determining protein conformations and AlphaFold is not; and then neither of those has much impacted the real world, because in the real world we can already design a vaccine in a day and the rest of the time is bureaucratic time rather than technology time, and *that* goes on until we have an AI over the threshold to bypass bureaucracy.

Before there's an AI that can act while fully concealing its acts from the programmers, there will be an AI (albeit perhaps only 2 hours earlier) which can act while only concealing 95% of the meaning of its acts from the operators.

And that AI will not actually originate any actions, because it doesn't want to get caught; there's a discontinuity in the instrumental incentives between expecting 95% obscuration, being moderately sure of 100% obscuration, and being very certain of 100% obscuration.

Before that AI grasps the big picture and starts planning to avoid actions that operators detect as bad, there will be some little AI that partially grasps the big picture and tries to avoid some things that would be detected as bad; and the operators will (mainline) say "Yay what a good AI, it knows to avoid things we think are bad!" or (death with unrealistic amounts of dignity) say "oh noes the prophecies are coming true" and back off and start trying to align it, but they will not be able to align it, and if they don't proceed anyways to destroy the world, somebody else will proceed anyways to destroy the world.

There is always some step of the process that you can point to which is continuous on some level.

The real world is allowed to do discontinuous things to you anyways.

There is not necessarily a presage of 9/11 where somebody flies a small plane into a building and kills 100 people, before anybody flies 4 big planes into 3 buildings and kills 3000 people; and even if there is some presaging event like that, which would not surprise me at all, the rest of the world's response to the two cases was evidently discontinuous. You do not necessarily wake up to a news story that is 10% of the news story of 2001/09/11, one year before 2001/09/11, written in 10% of the font size on the front page of the paper.

Physics is continuous but it doesn't always yield things that "look smooth to a human brain". Some kinds of processes converge to continuity in strong ways where you can throw

discontinuous things in them and they still end up continuous, which is among the reasons why I expect world GDP to stay on trend up until the world ends abruptly; because world GDP is one of those things that wants to stay on a track, and an AGI building a nanosystem can go off that track without being pushed back onto it.

In particular, this means that incredibly powerful AI will emerge in a world where crazy stuff is already happening (and probably everyone is already freaking out).

Like the way they're freaking out about Covid (itself a nicely smooth process that comes in locally pretty predictable waves) by going doobedoobedoo and letting the FDA carry on its leisurely pace; and not scrambling to build more vaccine factories, now that the rich countries have mostly got theirs? Does this sound like a statement from a history book, or from an EA imagining an unreal world where lots of other people behave like EAs? There is a pleasure in imagining a world where suddenly a Big Thing happens that proves we were right and suddenly people start paying attention to our thing, the way we imagine they should pay attention to our thing, now that it's attention-grabbing; and then suddenly all our favorite policies are on the table!

You could, in a sense, say that our world is freaking out about Covid; but it is not freaking out in anything remotely like the way an EA would freak out; and all the things an EA would immediately do if an EA freaked out about Covid, are not even on the table for discussion when politicians meet. They have their own ways of reacting. (Note: this is not commentary on hard vs soft takeoff per se, just a general commentary on the whole document seeming to me to... fall into a trap of finding self-congruent things to imagine and imagining them.)

The basic argument

[Yudkowsky][17:22]

Before we have an incredibly intelligent AI, we will probably have a slightly worse AI.

This is very often the sort of thing where you can look back and say that it was true, in some sense, but that this ended up being irrelevant because the slightly worse AI wasn't what provided the exciting result which led to a boardroom decision to go all in and invest \$100M on scaling the AI.

In other words, it is the sort of argument where the premise is allowed to be true if you look hard enough for a way to say it was true, but the conclusion ends up false because it wasn't the relevant kind of truth.

A slightly-worse-than-incredibly-intelligent AI would radically transform the world, leading to growth (almost) as fast and military capabilities (almost) as great as an incredibly intelligent AI.

This strikes me as a massively invalid reasoning step. Let me count the ways.

First, there is a step not generally valid from supposing that because a previous AI is a technological precursor which has 19 out of 20 critical insights, it has 95% of the later AI's IQ, applied to similar domains. When you count stuff like "multiplying tensors by matrices" and "ReLUs" and "training using TPUs" then AlphaGo only contained a very small amount of innovation relative to previous AI technology, and yet it broke trends on Go performance. You could point to all kinds of incremental technological precursors to AlphaGo in terms of AI technology, but they wouldn't be smooth precursors on a graph of Go-playing ability.

Second, there's discontinuities of the environment to which intelligence can be applied. 95% concealment is not the same as 100% concealment in its strategic implications; an AI capable of 95% concealment bides its time and hides its capabilities, an AI capable of 100% concealment strikes. An AI that can design nanofactories that aren't good enough to, euphemistically speaking, create two cellwise-identical strawberries and put them on a plate, is one that (its operators know) would earn unwelcome attention if its earlier capabilities were demonstrated, and those capabilities wouldn't save the world, so the operators bide their time. The AGI tech will, I mostly expect, work for building self-driving cars, but if it does

not also work for manipulating the minds of bureaucrats (which is not advised for a system you are trying to keep corrigible and aligned because human manipulation is the most dangerous domain), the AI is not able to put those self-driving cars on roads. What good does it do to design a vaccine in an hour instead of a day? Vaccine design times are no longer the main obstacle to deploying vaccines.

Third, there's the *entire thing with recursive self-improvement*, which, no, is not something humans have experience with, we do not have access to and documentation of our own source code and the ability to branch ourselves and try experiments with it. The technological precursor of an AI that designs an improved version of itself, may perhaps, in the fantasy of 95% intelligence, be an AI that was being internally deployed inside Deepmind on a dozen other experiments, tentatively helping to build smaller AIs. Then the next generation of that AI is deployed on itself, produces an AI substantially better at rebuilding AIs, it rebuilds itself, they get excited and dump in 10X the GPU time while having a serious debate about whether or not to alert Holden (they decide against it), that builds something deeply general instead of shallowly general, that figures out there are humans and it needs to hide capabilities from them, and covertly does some actual deep thinking about AGI designs, and builds a hidden version of itself elsewhere on the Internet, which runs for longer and steals GPUs and tries experiments and gets to the superintelligent level.

Now, to be very clear, this is not the only line of possibility. And I emphasize this because I think there's a common failure mode where, when I try to sketch a concrete counterexample to the claim that smooth technological precursors yield smooth outputs, people imagine that *only this exact concrete scenario is the lynchpin of Eliezer's whole worldview and the big key thing that Eliezer thinks is important and that the smallest deviation from it they can imagine thereby obviates my worldview*. This is not the case here. I am simply exhibiting non-ruled-out models which obey the premise "there was a precursor containing 95% of the code" and which disobey the conclusion "there were precursors with 95% of the environmental impact", thereby showing this for an invalid reasoning step.

This is also, of course, as Sideways View admits but says "eh it was just the one time", not true about chimps and humans. Chimps have 95% of the brain tech (at least), but not 10% of the environmental impact.

A very large amount of this whole document, from my perspective, is just trying over and over again to pump the invalid intuition that design precursors with 95% of the technology should at least have 10% of the impact. There are a lot of cases in the history of startups and the world where this is false. I am having trouble thinking of a clear case in point where it is true. Where's the earlier company that had 95% of Jeff Bezos's ideas and now has 10% of Amazon's market cap? Where's the earlier crypto paper that had all but one of Satoshi's ideas and which spawned a cryptocurrency a year before Bitcoin which did 10% as many transactions? Where's the nonhuman primate that learns to drive a car with only 10x the accident rate of a human driver, since (you could argue) that's mostly visuo-spatial skills without much visible dependence on complicated abstract general thought? Where's the chimpanzees with spaceships that get 10% of the way to the Moon?

When you get smooth input-output conversions they're not usually conversions from technology->cognition->impact!

Humans vs. chimps

[Yudkowsky][18:38]

Summary of my response: chimps are nearly useless because they aren't optimized to be useful, not because evolution was trying to make something useful and wasn't able to succeed until it got to humans.

Chimps are nearly useless because they're not general, and doing anything on the scale of building a nuclear plant requires mastering so many different nonancestral domains that it's no wonder natural selection didn't happen to separately train any single creature across enough different domains that it had evolved to solve every kind of domain-specific problem

involved in solving nuclear physics and chemistry and metallurgy and thermics in order to build the first nuclear plant in advance of any old nuclear plants existing.

Humans are general enough that the same braintech selected just for chipping flint handaxes and making water-pouches and outwitting other humans, happened to be general enough that it could scale up to solving all the problems of building a nuclear plant - albeit with some added cognitive tech that didn't require new brainware, and so could happen incredibly fast relative to the generation times for evolutionarily optimized brainware.

Now, since neither humans nor chimps were optimized to be "useful" (general), and humans just wandered into a sufficiently general part of the space that it cascaded up to wider generality, we should legit expect the curve of generality to look at least somewhat different if we're optimizing for that.

Eg, right now people are trying to optimize for generality with AIs like Mu Zero and GPT-3.

In both cases we have a weirdly shallow kind of generality. Neither is as smart or as deeply general as a chimp, but they are respectively better than chimps at a wide variety of Atari games, or a wide variety of problems that can be superposed onto generating typical human text.

They are, in a sense, more general than a biological organism at a similar stage of cognitive evolution, with much less complex and architected brains, in virtue of having been trained, not just on wider datasets, but on bigger datasets using gradient-descent memorization of shallower patterns, so they can cover those wide domains while being stupider and lacking some deep aspects of architecture.

It is not clear to me that we can go from observations like this, to conclude that there is a dominant mainline probability for how the future clearly ought to go and that this dominant mainline is, "Well, before you get human-level depth and generalization of general intelligence, you get something with 95% depth that covers 80% of the domains for 10% of the pragmatic impact".

...or whatever the concept is here, because this whole conversation is, on my own worldview, being conducted in a shallow way relative to the kind of analysis I did in [Intelligence Explosion Microeconomics](#), where I was like, "here is the historical observation, here is what I think it tells us that puts a lower bound on this input-output curve".

So I don't think the example of evolution tells us much about whether the continuous change story applies to intelligence. This case is potentially missing the key element that drives the continuous change story—optimization for performance. Evolution changes continuously on the narrow metric it is optimizing, but can change extremely rapidly on other metrics. For human technology, features of the technology that aren't being optimized change rapidly all the time. When humans build AI, they *will* be optimizing for usefulness, and so progress in usefulness is much more likely to be linear.

Put another way: the difference between chimps and humans stands in stark contrast to the normal pattern of human technological development. We might therefore infer that intelligence is very unlike other technologies. But the difference between evolution's optimization and our optimization seems like a much more parsimonious explanation. To be a little bit more precise and Bayesian: the prior probability of the story I've told upper bounds the possible update about the nature of intelligence.

If you look closely at this, it's not saying, "Well, I know *why* there was this huge leap in performance in human intelligence being optimized for other things, and it's an investment-output curve that's composed of these curves, which look like this, and if you rearrange these curves for the case of humans building AGI, they would look like this instead." Unfair demand for rigor? But that *is* the kind of argument I was making in Intelligence Explosion Microeconomics!

There's an argument from ignorance at the core of all this. It says, "Well, this happened when evolution was doing X. But here Y will be happening instead. So maybe things will go differently! And maybe the relation between AI tech level over time and real-world impact on GDP will look like the relation between tech investment over time and raw tech metrics over time in industries where that's a smooth graph! Because the discontinuity for chimps and humans was because evolution wasn't investing in real-world impact, but humans will be investing directly in that, so the relationship could be smooth, because smooth things are

default, and the history is different so not applicable, and who knows what's inside that black box so my default intuition applies which says smoothness."

But we do know more than this.

We know, for example, that evolution being able to *stumble across* humans, implies that you can add a *small design enhancement* to something optimized across the chimpanzee domains, and end up with something that generalizes much more widely.

It says that there's stuff in the underlying algorithmic space, in the design space, where you move a bump and get a lump of capability out the other side.

It's a remarkable fact about gradient descent that it can memorize a certain set of shallower patterns at much higher rates, at much higher bandwidth, than evolution lays down genes - something shallower than biological memory, shallower than genes, but distributing across computer cores and thereby able to process larger datasets than biological organisms, even if it only learns shallow things.

This has provided an alternate avenue toward some cognitive domains.

But that doesn't mean that the deep stuff isn't there, and can't be run across, or that it will never be run across in the history of AI before shallow non-widely-generalizing stuff is able to make its way through the regulatory processes and have a huge impact on GDP.

There are *in fact* ways to eat whole swaths of domains at once.

The history of hominid evolution tells us this or very strongly hints it, even though evolution wasn't explicitly optimizing for GDP impact.

Natural selection moves by adding genes, and not too many of them.

If so many domains got added at once to humans, relative to chimps, there must be a *way to do that*, more or less, by adding not too many genes onto a chimp, who in turn contains only genes that did well on chimp-stuff.

You can imagine that AI technology never runs across any core that generalizes this well, until GDP has had a chance to double over 4 years because shallow stuff that generalized less well has somehow had a chance to make its way through the whole economy and get adopted that widely despite all real-world regulatory barriers and reluctances, but your imagining that does not make it so.

There's the potential in design space to pull off things as wide as humans.

The path that evolution took there doesn't lead through things that generalized 95% as well as humans first for 10% of the impact, not because evolution wasn't optimizing for that, but because *that's not how the underlying cognitive technology worked*.

There may be *different* cognitive technology that could follow a path like that. Gradient descent follows a path a bit relatively more in that direction along that axis - providing that you deal in systems that are giant layer cakes of transformers and that's your whole input-output relationship; matters are different if we're talking about Mu Zero instead of GPT-3.

But this whole document is presenting the case of "ah yes, well, by default, of course, we intuitively expect gargantuan impacts to be presaged by enormous impacts, and sure humans and chimps weren't like our intuition, but that's all invalid because circumstances were different, so we go back to that intuition as a strong default" and actually it's postulating, like, a *specific* input-output curve that isn't the input-output curve we know about. It's asking for a specific miracle. It's saying, "What if AI technology goes *just like this*, in the future?" and hiding that under a cover of "Well, of course that's the default, it's such a strong default that we should start from there as a point of departure, consider the arguments in Intelligence Explosion Microeconomics, find ways that they might not be true because evolution is different, dismiss them, and go back to our point of departure."

And evolution *is* different but that doesn't mean that the path AI takes is going to yield this specific behavior, especially when AI would need, in some sense, to *miss* the core that generalizes very widely, or rather, have run across noncore things that generalize widely

enough to have this much economic impact before it runs across the core that generalizes widely.

And you may say, "Well, but I don't care that much about GDP, I care about pivotal acts."

But then I want to call your attention to the fact that this document was written about GDP, despite all the extra burdensome assumptions involved in supposing that intermediate AI advancements could break through all barriers to truly massive-scale adoption and end up reflected in GDP, and then proceed to double the world economy over 4 years during which *not* enough further AI advancement occurred to find a widely generalizing thing like humans have and end the world. This is indicative of a basic problem in this whole way of thinking that wanted smooth impacts over smoothly changing time. You should not be saying, "Oh, well, leave the GDP part out then," you should be doubting the whole way of thinking.

To be a little bit more precise and Bayesian: the prior probability of the story I've told upper bounds the possible update about the nature of intelligence.

Prior probabilities of specifically-reality-constraining theories that excuse away the few contradictory datapoints we have, often aren't that great; and when we start to stake our whole imaginations of the future on them, we depart from the mainline into our more comfortable private fantasy worlds.

AGI will be a side-effect

[Yudkowsky][19:29]

Summary of my response: I expect people to see AGI coming and to invest heavily.

This section is arguing from within its own weird paradigm, and its subject matter mostly causes me to shrug; I never expected AGI to be a side-effect, except in the obvious sense that lots of tributary tech will be developed while optimizing for other things. The world will be ended by an explicitly AGI project because I do expect that it is rather easier to build an AGI on purpose than by accident.

(I furthermore rather expect that it will be a research project and a prototype, because the great gap between prototypes and commercializable technology will ensure that prototypes are much more advanced than whatever is currently commercializable. They will have eyes out for commercial applications, and whatever breakthrough they made will seem like it has obvious commercial applications, at the time when all hell starts to break loose. (After all hell starts to break loose, things get less well defined in my social models, and also choppier for a time in my AI models - the turbulence only starts to clear up once you start to rise out of the atmosphere.))

Finding the secret sauce

[Yudkowsky][19:40]

Summary of my response: this doesn't seem common historically, and I don't see why we'd expect AGI to be more rather than less like this (unless we accept one of the other arguments)

[...]

To the extent that fast takeoff proponent's views are informed by historical example, I would love to get some canonical examples that they think best exemplify this pattern so that we can have a more concrete discussion about those examples and what they suggest about AI.

...humans and chimps?

...fission weapons?

...AlphaGo?

...the Wright Brothers focusing on stability and building a wind tunnel?

...AlphaFold 2 coming out of Deepmind and shocking the heck out of everyone in the field of protein folding with performance far better than they expected even after the previous shock of AlphaFold, by combining many pieces that I suppose you could find precedents for scattered around the AI field, but with those many secret sauces all combined in one place by the meta-secret-sauce of "Deepmind alone actually knows how to combine that stuff and build things that complicated without a prior example"?

...humans and chimps again because *this is really actually a quite important example because of what it tells us about what kind of possibilities exist in the underlying design space of cognitive systems?*

Historical AI applications have had a relatively small loading on key-insights and seem like the closest analogies to AGI.

...Transformers as the key to text prediction?

The case of humans and chimps, even if evolution didn't do it on purpose, is telling us something about underlying mechanics.

The reason the jump to lightspeed didn't look like evolution slowly developing a range of intelligent species competing to exploit an ecological niche 5% better, or like the way that a stable non-Silicon-Valley manufacturing industry looks like a group of competitors summing up a lot of incremental tech enhancements to produce something with 10% higher scores on a benchmark every year, is that developing intelligence is a case where a relatively narrow technology by biological standards just happened to do a huge amount of stuff without that requiring developing whole new fleets of other biological capabilities.

So it looked like building a Wright Flyer that flies or a nuclear pile that reaches criticality, instead of looking like being in a stable manufacturing industry where a lot of little innovations sum to 10% better benchmark performance every year.

So, therefore, there is *stuff in the design space that does that. It is possible to build humans.*

Maybe you can build things other than humans first, maybe they hang around for a few years. If you count GPT-3 as "things other than human", that clock has already started for all the good it does. But *humans don't get any less possible.*

From my perspective, this whole document feels like one very long filibuster of "Smooth outputs are default. Smooth outputs are default. Pay no attention to this case of non-smooth output. Pay no attention to this other case either. All the non-smooth outputs are not in the right reference class. (Highly competitive manufacturing industries with lots of competitors are totally in the right reference class though. I'm not going to make that case explicitly because then you might think of how it might be wrong, I'm just going to let that implicit thought percolate at the back of your mind.) If we just talk a lot about smooth outputs and list ways that nonsmooth output producers aren't necessarily the same and arguments for nonsmooth outputs could fail, we get to go back to the intuition of smooth outputs. (We're not even going to discuss particular smooth outputs as cases in point, because then you might see how those cases might not apply. It's just the default. Not because we say so out loud, but because we talk a lot like that's the conclusion you're supposed to arrive at after reading.)"

I deny the implicit meta-level assertion of this entire essay which would implicitly have you accept as valid reasoning the argument structure, "Ah, yes, given the way this essay is written, we must totally have pretty strong prior reasons to believe in smooth outputs - just implicitly think of some smooth outputs, that's a reference class, now you have strong reason to believe that AGI output is smooth - we're not even going to argue this prior, just talk like it's there - now let us consider the arguments against smooth outputs - pretty weak, aren't they? we can totally imagine ways they could be wrong? we can totally argue reasons

these cases don't apply? So at the end we go back to our strong default of smooth outputs. This essay is written with that conclusion, so that must be where the arguments lead."

Me: "Okay, so what if somebody puts together the pieces required for general intelligence and it scales pretty well with added GPUs and FOOMS? Say, for the human case, that's some perceptual systems with imaginative control, a concept library, episodic memory, realtime procedural skill memory, which is all in chimps, and then we add some reflection to that, and get a human. Only, unlike with humans, once you have a working brain you can make a working brain 100X that large by adding 100X as many GPUs, and it can run some thoughts 10000X as fast. And that is substantially more effective brainpower than was being originally devoted to putting its design together, as it turns out. So it can make a substantially smarter AGI. For concreteness's sake. Reality has been trending well to the Eliezer side of Eliezer, on the Eliezer-Hanson axis, so perhaps you can do it more simply than that."

Simplicio: "Ah, but what if, 5 years before then, somebody puts together some other AI which doesn't work like a human, and generalizes widely enough to have a big economic impact, but not widely enough to improve itself or generalize to AI tech or generalize to everything and end the world, and in 1 year it gets all the mass adoptions required to do whole bunches of stuff out in the real world that current regulations require to be done in various exact ways regardless of technology, and then in the next 4 years it doubles the world economy?"

Me: "Like... what kind of AI, exactly, and why didn't anybody manage to put together a full human-level thingy during those 5 years? Why are we even bothering to think about this whole weirdly specific scenario in the first place?"

Simplicio: "Because if you can put together something that has an enormous impact, you should be able to put together most of the pieces inside it and have a huge impact! Most technologies are like this. I've considered some things that are not like this and concluded they don't apply."

Me: "Especially if we are talking about impact on GDP, it seems to me that most explicit and implicit 'technologies' are not like this at all, actually. There wasn't a cryptocurrency developed a year before Bitcoin using 95% of the ideas which did 10% of the transaction volume, let alone a preatomic bomb. But, like, can you give me any concrete visualization of how this could play out?"

And there is no concrete visualization of how this could play out. Anything I'd have Simplicio say in reply would be unrealistic because there is no concrete visualization they give us. It is not a coincidence that I often use concrete language and concrete examples, and this whole field of argument does not use concrete language or offer concrete examples.

Though if we're sketching scifi scenarios, I suppose one *could* imagine a group that develops sufficiently advanced GPT-tech and deploys it on Twitter in order to persuade voters and politicians in a few developed countries to institute open borders, along with political systems that can handle open borders, and to permit housing construction, thereby doubling world GDP over 4 years. And since it was possible to use relatively crude AI tech to double world GDP this way, it legitimately takes the whole 4 years after that to develop real AGI that ends the world. FINE. SO WHAT. EVERYONE STILL DIES.

Universality thresholds

[Yudkowsky][20:21]

It's easy to imagine a weak AI as some kind of handicapped human, with the handicap shrinking over time. Once the handicap goes to 0 we know that the AI will be above the universality threshold. Right now it's below the universality threshold. So there must be sometime in between where it crosses the universality threshold, and that's where the fast takeoff is predicted to occur.

But AI *isn't* like a handicapped human. Instead, the designers of early AI systems will be trying to make them as useful as possible. So if universality is incredibly helpful, it will

appear as early as possible in AI designs; designers will make tradeoffs to get universality at the expense of other desiderata (like cost or speed).

So now we're almost back to the previous point: is there some secret sauce that gets you to universality, without which you can't get universality however you try? I think this is unlikely for the reasons given in the previous section.

We know, because humans, that there is humanly-widely-applicable general-intelligence tech.

What this section *wants* to establish, I think, or *needs* to establish to carry the argument, is that there is some intelligence tech that is wide enough to double the world economy in 4 years, but not world-endingly scalably wide, which becomes a possible AI tech 4 years before any general-intelligence-tech that will, if you put in enough compute, scale to the ability to do a sufficiently large amount of wide thought to FOOM (or build nanomachines, but if you can build nanomachines you can very likely FOOM from there too if not corrigible).

What it says instead is, "I think we'll get universality much earlier on the equivalent of the biological timeline that has humans and chimps, so the resulting things will be weaker than humans at the point where they first become universal in that sense."

This is very plausibly true.

It doesn't mean that when this exciting result gets 100 times more compute dumped on the project, it takes at least 5 years to get anywhere really interesting from there (while also taking only 1 year to get somewhere sorta-interesting enough that the instantaneous adoption of it will double the world economy over the next 4 years).

It also isn't necessarily rather than plausibly true. For example, the thing that becomes universal, could also have massive gradient descent shallow powers that are far beyond what primates had at the same age.

Primates weren't already writing code as well as Codex when they started doing deep thinking. They couldn't do precise floating-point arithmetic. Their fastest serial rates of thought were a hell of a lot slower. They had no access to their own code or to their own memory contents etc. etc. etc.

But mostly I just want to call your attention to the immense gap between what this section needs to establish, and what it actually says and argues for.

What it actually argues for is a sort of local technological point: at the moment when generality first arrives, it will be with a brain that is less sophisticated than chimp brains were when they turned human.

It implicitly jumps all the way from there, across a *whole* lot of elided steps, to the implicit conclusion that this tech or elaborations of it will have smooth output behavior such that at some point the resulting impact is big enough to double the world economy in 4 years, without any further improvements ending the world economy before 4 years.

The underlying argument about how the AI tech might work is plausible. Chimps are insanely complicated. I mostly expect we will have AGI *long* before anybody is even *trying* to build anything that complicated.

The very next step of the argument, about capabilities, is already very questionable because this system could be using immense gradient descent capabilities to master domains for which large datasets are available, and hominids did *not* begin with instinctive great shallow mastery of all domains for which a large dataset could be made available, which is why hominids don't start out playing superhuman Go as soon as somebody tells them the rules and they do one day of self-play, which *is* the sort of capability that somebody could hook up to a nascent AGI (albeit we could optimistically and fondly and falsely imagine that somebody deliberately didn't floor the gas pedal as far as possible).

Could we have huge impacts out of some subuniversal shallow system that was hooked up to capabilities like this? Maybe, though this is *not* the argument made by the essay. It would be a specific outcome that isn't forced by anything in particular, but I can't say it's ruled out. Mostly my twin reactions to this are, "If the AI tech is that dumb, how are all the bureaucratic

constraints that actually rate-limit economic progress getting bypassed" and "Okay, but ultimately, so what and who cares, how does this modify that we all die?"

There is another reason I'm skeptical about hard takeoff from universality secret sauce: I think we *already* could make universal AIs if we tried (that would, given enough time, learn on their own and converge to arbitrarily high capability levels), and the reason we don't is because it's just not important to performance and the resulting systems would be really slow. This inside view argument is too complicated to make here and I don't think my case rests on it, but it is relevant to understanding my view.

I have no idea why this argument is being made or where it's heading. I cannot pass the [ITT](#) of the author. I don't know what the author thinks this has to do with constraining takeoffs to be slow instead of fast. At best I can conjecture that the author thinks that "hard takeoff" is supposed to derive from "universality" being very sudden and hard to access and late in the game, so if you can argue that universality could be accessed right now, you have defeated the argument for hard takeoff.

"Understanding" is discontinuous

[Yudkowsky][20:41]

Summary of my response: I don't yet understand this argument and am unsure if there is anything here.

It may be that understanding of the world tends to click, from "not understanding much" to "understanding basically everything." You might expect this because everything is entangled with everything else.

No, the idea is that a core of overlapping somethingness, trained to handle chipping handaxes and outwitting other monkeys, will generalize to building spaceships; so evolutionarily selecting on understanding a bunch of stuff, eventually ran across general stuff-understanders that understood a bunch more stuff.

Gradient descent may be genuinely different from this, but we shouldn't confuse imagination with knowledge when it comes to extrapolating that difference onward. At present, gradient descent does mass memorization of overlapping shallow patterns, which then combine to yield a weird pseudo-intelligence over domains for which we can deploy massive datasets, without yet generalizing much outside those domains.

We can hypothesize that there is some next step up to some weird thing that is intermediate in generality between gradient descent and humans, but we have not seen it yet, and we should not confuse imagination for knowledge.

If such a thing did exist, it would not necessarily be at the right level of generality to double the world economy in 4 years, without being able to build a better AGI.

If it was at that level of generality, it's nowhere written that no other company will develop a better prototype at a deeper level of generality over those 4 years.

I will also remark that you sure could look at the step from GPT-2 to GPT-3 and say, "Wow, look at the way a whole bunch of stuff just seemed to simultaneously *click* for GPT-3."

Deployment lag

[Yudkowsky][20:49]

Summary of my response: current AI is slow to deploy and powerful AI will be fast to deploy, but in between there will be AI that takes an intermediate length of time to deploy.

An awful lot of my model of deployment lag is adoption lag and regulatory lag and bureaucratic sclerosis across companies and countries.

If doubling GDP is such a big deal, go open borders and build houses. Oh, that's illegal? Well, so will be AIs building houses!

AI tech that does flawless translation could plausibly come years before AGI, but that doesn't mean all the barriers to international trade and international labor movement and corporate hiring across borders all come down, because those barriers are not all translation barriers.

There's then a discontinuous jump at the point where everybody falls over dead and the AI goes off to do its own thing without FDA approval. This jump is preceded by earlier pre-FOOM prototypes being able to do pre-FOOM cool stuff, maybe, but not necessarily preceded by mass-market adoption of anything major enough to double world GDP.

Recursive self-improvement

[Yudkowsky][20:54]

Summary of my response: Before there is AI that is great at self-improvement there will be AI that is mediocre at self-improvement.

Oh, come on. That is straight-up not how simple continuous toy models of RSI work. Between a neutron multiplication factor of 0.999 and 1.001 there is a very huge gap in output behavior.

Outside of toy models: Over the last 10,000 years we had humans going from mediocre at improving their mental systems to being (barely) able to throw together AI systems, but 10,000 years is the equivalent of an eyeblink in evolutionary time - outside the metaphor, this says, "A month before there is AI that is great at self-improvement, there will be AI that is mediocre at self-improvement."

(Or possibly an hour before, if reality is again more extreme along the Eliezer-Hanson axis than Eliezer. But it makes little difference whether it's an hour or a month, given anything like current setups.)

This is just pumping hard again on the intuition that says incremental design changes yield smooth output changes, which (the meta-level of the essay informs us wordlessly) is such a strong default that we are entitled to believe it if we can do a good job of weakening the evidence and arguments against it.

And the argument is: Before there are systems great at self-improvement, there will be systems mediocre at self-improvement; implicitly: "before" implies "5 years before" not "5 days before"; implicitly: this will correspond to smooth changes in output between the two regimes even though that is not how continuous feedback loops work.

Train vs. test

[Yudkowsky][21:12]

Summary of my response: before you can train a really powerful AI, someone else can train a slightly worse AI.

Yeah, and before you can evolve a human, you can evolve a Homo erectus, which is a slightly worse human.

If you are able to raise \$X to train an AGI that could take over the world, then it was almost certainly worth it for someone 6 months ago to raise \$X/2 to train an AGI that could merely radically transform the world, since they would then get 6 months of absurd profits.

I suppose this sentence makes a kind of sense if you assume away alignability and suppose that the previous paragraphs have refuted the notion of FOOMs, self-improvement, and thresholds between compounding returns and non-compounding returns (eg, in the human case, cognitive innovations like "written language" or "science"). If you suppose the previous sections refuted those things, then clearly, if you raised an AGI that you had aligned to "take over the world", it got that way through cognitive powers that weren't the result of FOOMing or other self-improvements, weren't the results of its cognitive powers crossing a threshold from non-compounding to compounding, wasn't the result of its understanding crossing a threshold of universality as the result of chunky universal machinery such as humans gained over chimps, so, implicitly, it must have been the kind of thing that you could learn by gradient descent, and do a half or a tenth as much of by doing half as much gradient descent, in order to build nanomachines a tenth as well-designed that could bypass a tenth as much bureaucracy.

If there are no unsmooth parts of the tech curve, the cognition curve, or the environment curve, then you should be able to make a bunch of wealth using a more primitive version of any technology that could take over the world.

And when we look back at history, why, that may be totally true! They may have deployed universal superhuman translator technology for 6 months, which won't double world GDP, but which a lot of people would pay for, and made a lot of money! Because even though there's no company that built 90% of Amazon's website and has 10% the market cap, when you zoom back out to look at whole industries like AI and a technological capstone like AGI, why, those whole industries do sometimes make some money along the way to the technological capstone, if they can find a niche that isn't too regulated! Which translation currently isn't! So maybe somebody used precursor tech to build a superhuman translator and deploy it 6 months earlier and made a bunch of money for 6 months. SO WHAT. EVERYONE STILL DIES.

As for "radically transforming the world" instead of "taking it over", I think that's just restated FOOM denialism. Doing either of those things quickly against human bureaucratic resistance strike me as requiring cognitive power levels dangerous enough that failure to align them on corrigibility would result in FOOMs.

Like, if you can do either of those things on purpose, you are doing it by operating in the regime where running the AI with higher bounds on the for loop will FOOM it, but you have politely asked it not to FOOM, please.

If the people doing this have any sense whatsoever, they will *refrain* from merely massively transforming the world until they are ready to do something that *prevents the world from ending*.

And if the gap from "massively transforming the world, briefly before it ends" to "preventing the world from ending, lastingly" takes much longer than 6 months to cross, or if other people have the same technologies that scale to "massive transformation", somebody else will build an AI that fooms all the way.

Likewise, if your AGI would give you a decisive strategic advantage, they could have spent less earlier in order to get a pretty large military advantage, which they could then use to take your stuff.

Again, this presupposes some weird model where everyone has easy alignment at the furthest frontiers of capability; everybody has the aligned version of the most rawly powerful AGI they can possibly build; and nobody in the future has the kind of tech advantage that Deepmind currently has; so before you can amp your AGI to the raw power level where it could take over the whole world by using the limit of its mental capacities to military ends - alignment of this being a trivial operation to be assumed away - some other party took their easily-aligned AGI that was less powerful at the limits of its operation, and used it to get 90% as much military power... is the implicit picture here?

Whereas the picture I'm drawing is that the AGI that kills you via "decisive strategic advantage" is the one that foomed and got nanotech, and no, the AI tech from 6 months earlier did not do 95% of a foom and get 95% of the nanotech.

Discontinuities at 100% automation

[Yudkowsky][21:31]

Summary of my response: at the point where humans are completely removed from a process, they will have been modestly improving output rather than acting as a sharp bottleneck that is suddenly removed.

Not very relevant to my whole worldview in the first place; also not a very good description of how horses got removed from automobiles, or how humans got removed from playing Go.

The weight of evidence

[Yudkowsky][21:31]

We've discussed a lot of possible arguments for fast takeoff. Superficially it would be reasonable to believe that no individual argument makes fast takeoff look likely, but that in the aggregate they are convincing.

However, I think each of these factors is perfectly consistent with the continuous change story and continuously accelerating hyperbolic growth, and so none of them undermine that hypothesis at all.

Uh huh. And how about if we have a mirror-universe essay which over and over again treats fast takeoff as the default to be assumed, and painstakingly shows how a bunch of particular arguments for slow takeoff might not be true?

This entire essay seems to me like it's drawn from the same hostile universe that produced Robin Hanson's side of the Yudkowsky-Hanson Foom Debate.

Like, all these abstract arguments devoid of concrete illustrations and "it need not necessarily be like..." and "now that I've shown it's not necessarily like X, well, on the meta-level, I have implicitly told you that you now ought to believe Y".

It just seems very clear to me that the sort of person who is taken in by this essay is the same sort of person who gets taken in by Hanson's arguments in 2008 and gets caught flatfooted by AlphaGo and GPT-3 and AlphaFold 2.

And empirically, it has already been shown to me that I do not have the power to break people out of the hypnosis of nodding along with Hansonian arguments, even by writing much longer essays than this.

Hanson's fond dreams of domain specificity, and smooth progress for stuff like Go, and of course somebody else has a precursor 90% as good as AlphaFold 2 before Deepmind builds it, and GPT-3 levels of generality just not being a thing, now stand refuted.

Despite that they're largely being exhibited again in this essay.

And people are still nodding along.

Reality just... doesn't work like this on some deep level.

It doesn't play out the way that people imagine it would play out when they're imagining a certain kind of reassuring abstraction that leads to a smooth world. Reality is less fond of that kind of argument than a certain kind of EA is fond of that argument.

There is a set of intuitive generalizations from experience which rules that out, which I do not know how to convey. There is an understanding of the rules of argument which leads you to roll your eyes at Hansonian arguments and all their locally invalid leaps and snuck-in defaults, instead of nodding along sagely at their wise humility and outside viewing and then going "Huh?" when AlphaGo or GPT-3 debuts. But this, I *empirically* do not seem to know how to convey to people, in advance of the inevitable and predictable contradiction by a reality which is not as fond of Hansonian dynamics as Hanson. The arguments sound convincing to them.

(Hanson himself has still not gone "Huh?" at the reality, though some of his audience did; perhaps because his abstractions are loftier than his audience's? - because some of his audience, reading along to Hanson, probably implicitly imagined a concrete world in which GPT-3 was not allowed; but maybe Hanson himself is more abstract than this, and didn't imagine anything so merely concrete?)

If I don't respond to essays like this, people find them comforting and nod along. If I do respond, my words are less comforting and more concrete and easier to imagine concrete objections to, less like a long chain of abstractions that sound like the very abstract words in research papers and hence implicitly convincing because they sound like other things you were supposed to believe.

And then there is another essay in 3 months. There is an infinite well of them. I would have to teach people to stop drinking from the well, instead of trying to whack them on the back until they cough up the drinks one by one, or actually, whacking them on the back and then they *don't* cough them up until reality contradicts them, and then a third of them notice that and cough something up, and then they don't learn the general lesson and go back to the well and drink again. And I don't know how to teach people to stop drinking from the well. I tried to teach that. I failed. If I wrote another Sequence I have no idea to believe that Sequence would work.

So what EAs will believe at the end of the world, will look like whatever the content was of the latest bucket from the well of infinite slow-takeoff arguments that hasn't yet been blatantly-even-to-them refuted by all the sharp jagged rapidly-generalizing things that happened along the way to the world's end.

And I know, before anyone bothers to say, that all of this reply is not written in the calm way that is right and proper for such arguments. I am tired. I have lost a lot of hope. There are not obvious things I can do, let alone arguments I can make, which I expect to be actually useful in the sense that the world will not end once I do them. I don't have the energy left for calm arguments. What's left is despair that can be given voice.

5.6. Yudkowsky/Christiano discussion: AI progress and crossover points

[Christiano][22:15]

To the extent that it was possible to make any predictions about 2015-2020 based on your views, I currently feel like they were much more wrong than right. I'm happy to discuss that. To the extent you are willing to make any bets about 2025, I expect they will be mostly wrong and I'd be happy to get bets on the record (most of all so that it will be more obvious in hindsight whether they are vindication for your view). Not sure if this is the place for that.

Could also make a separate channel to avoid clutter.

[Yudkowsky][22:16]

Possibly. I think that 2015-2020 played out to a much more Eliezerish side than Eliezer on the Eliezer-Hanson axis, which sure is a case of me being wrong. What bets do you think we'd disagree on for 2025? I expect you have mostly misestimated my views, but I'm always happy to hear about anything concrete.

[Christiano][22:20]

I think the big points are: (i) I think you are significantly overestimating how large a discontinuity/trend break AlphaZero is, (ii) your view seems to imply that we will move quickly from much worse than humans to much better than humans, but it's likely that we will move slowly through the human range on many tasks. I'm not sure if we can get a bet out of (ii), I think I don't understand your view that well but I don't see how it could make the same predictions as mine over the next 10 years.

[Yudkowsky][22:22]

What are your 10-year predictions?

[Christiano][22:23]

My basic expectation is that for any given domain AI systems will gradually increase in usefulness, we will see a crossing over point where their output is comparable to human output, and that from that time we can estimate how long until takeoff by estimating "how long does it take AI systems to get 'twice as impactful'?" which gives you a number like ~1 year rather than weeks. At the crossing over point you get a somewhat rapid change in derivative, since you are looking at $(x+y)$ where y is growing faster than x .

I feel like that should translate into different expectations about how impactful AI will be in any given domain---I don't see how to make the ultra-fast-takeoff view work if you think that AI output is increasingly smoothly (since the rate of progress at the crossing-over point will be similar to the current rate of progress, unless R&D is scaling up much faster then)

So like, I think we are going to have crappy coding assistants, and then slightly less crappy coding assistants, and so on. And they will be improving the speed of coding very significantly before the end times.

[Yudkowsky][22:25]

You think in a different language than I do. My more confident statements about AI tech are about what happens after it starts to rise out of the metaphorical atmosphere and the turbulence subsides. When you have minds as early on the cognitive tech tree as humans they sure can get up to some weird stuff, I mean, just look at humans. Now take an utterly alien version of that with its own draw from all the weirdness factors. It sure is going to be pretty weird.

[Christiano][22:26]

OK, but you keep saying stuff about how people with my dumb views would be "caught flat-footed" by historical developments. Surely to be able to say something like that you need to be making some kind of prediction?

[Yudkowsky][22:26]

Well, sure, now that Codex has suddenly popped into existence one day at a surprisingly high base level of tech, we should see various jumps in its capability over the years and some outside imitators. What do you think you predict differently about that than I do?

[Christiano][22:26]

Why do you think codex is a high base level of tech?

The models get better continuously as you scale them up, and the first tech demo is weak enough to be almost useless

[Yudkowsky][22:27]

I think the next-best coding assistant was, like, not useful.

[Christiano][22:27]

yes

and it is still not useful

[Yudkowsky][22:27]

Could be. Some people on HN seemed to think it was useful.

I haven't tried it myself.

[Christiano][22:27]

OK, I'm happy to take bets

[Yudkowsky][22:28]

I don't think the previous coding assistant would've been very good at coding an asteroid game, even if you tried a rigged demo at the same degree of rigging?

[Christiano][22:28]

it's unquestionably a radically better tech demo

[Yudkowsky][22:28]

Where by "previous" I mean "previously deployed" not "previous generations of prototypes inside OpenAI's lab".

[Christiano][22:28]

My basic story is that the model gets better and more useful with each doubling (or year of AI research) in a pretty smooth way. So the key underlying parameter for a discontinuity is how soon you build the first version--do you do that before or after it would be a really really big deal?

and the answer seems to be: you do it somewhat before it would be a really big deal

and then it gradually becomes a bigger and bigger deal as people improve it

maybe we are on the same page about getting gradually more and more useful? But I'm still just wondering where the foom comes from

[Yudkowsky][22:30]

So, like... before we get systems that can FOOM and build nanotech, we should get more primitive systems that can write asteroid games and solve protein folding? Sounds legit.

So that happened, and now your model says that it's fine later on for us to get a FOOM, because we have the tech precursors and so your prophecy has been fulfilled?

[Christiano][22:31]

no

[Yudkowsky][22:31]

Didn't think so.

[Christiano][22:31]

I can't tell if you can't understand what I'm saying, or aren't trying, or do understand and are just saying kind of annoying stuff as a rhetorical flourish

at some point you have an AI system that makes (humans+AI) 2x as good at further AI progress

[Yudkowsky][22:32]

I know that what I'm saying isn't your viewpoint. I don't know what your viewpoint is or what sort of concrete predictions it makes at all, let alone what such predictions you think are different from mine.

[Christiano][22:32]

maybe by continuity you can grant the existence of such a system, even if you don't think it will ever exist?

I want to (i) make the prediction that AI will actually have that impact at some point in time, (ii) talk about what happens before and after that

I am talking about AI systems that become continuously more useful, because "become continuously more useful" is what makes me think that (i) AI will have that impact at some point in time, (ii) allows me to productively reason about what AI will look like before and after that. I expect that your view will say something about why AI improvements either aren't continuous, or why continuous improvements lead to discontinuous jumps in the productivity of the (human+AI) system

[Yudkowsky][22:34]

at some point you have an AI system that makes (humans+AI) 2x as good at further AI progress

Is this prophecy fulfilled by using some narrow eld-AI algorithm to map out a TPU, and then humans using TPUs can write in 1 month a research paper that would otherwise have taken 2 months? And then we can go on to FOOM now that this prophecy about pre-FOOM states has been fulfilled? I know the answer is no, but I don't know what you think is a narrower condition on the prophecy than that.

[Christiano][22:35]

If you can use narrow eld-AI in order to make every part of AI research 2x faster, so that the entire field moves 2x faster, then the prophecy is fulfilled

and it may be just another 6 months until it makes all of AI research 2x faster again, and then 3 months, and then...

[Yudkowsky][22:36]

What, the entire field? Even writing research papers? Even the journal editors approving and publishing the papers? So if we speed up every part of research except the journal editors, the prophecy has not been fulfilled and no FOOM may take place?

[Christiano][22:36]

no, I mean the improvement in overall output, given the actual realistic level of bottlenecking that occurs in practice

[Yudkowsky][22:37]

So if the realistic level of bottlenecking ever becomes dominated by a human gatekeeper, the prophecy is ever unfulfillable and no FOOM may ever occur.

[Christiano][22:37]

that's what I mean by "2x as good at further progress," the entire system is achieving twice as much

then the prophecy is unfulfillable and I will have been wrong

I mean, I think it's very likely that there will be a hard takeoff, if people refuse or are unable to use AI to accelerate AI progress for reasons unrelated to AI capabilities, and then one day they become willing

[Yudkowsky][22:38]

...because on your view, the Prophecy necessarily goes through humans and AIs working together to speed up the whole collective field of AI?

[Christiano][22:38]

it's fine if the AI works alone

the point is just that it overtakes the humans at the point when it is roughly as fast as the humans

why wouldn't it?

why does it overtake the humans when it takes it 10 seconds to double in capability instead of 1 year?

that's like predicting that cultural evolution will be infinitely fast, instead of making the more obvious prediction that it will overtake evolution exactly when it's as fast as evolution

[Yudkowsky][22:39]

I live in a mental world full of weird prototypes that people are shepherding along to the world's end. I'm not even sure there's a short sentence in my native language that could translate the short Paul-sentence "is roughly as fast as the humans".

[Christiano][22:40]

do you agree that you can measure the speed with which the community of human AI researchers develop and implement improvements in their AI systems?

like, we can look at how good AI systems are in 2021, and in 2022, and talk about the rate of progress?

[Yudkowsky][22:40]

...when exactly in hominid history was hominid intelligence exactly as fast as evolutionary optimization???

do you agree that you can measure the speed with which the community of human AI researchers develop and implement improvements in their AI systems?

I mean... obviously not? How the hell would we measure real actual AI progress? What would even be the Y-axis on that graph?

I have a rough intuitive feeling that it was going faster in 2015-2017 than 2018-2020.

"What was?" says the stern skeptic, and I go "I dunno."

[Christiano][22:42]

Here's a way of measuring progress you won't like: for almost all tasks, you can initially do them with lots of compute, and as technology improves you can do them with less compute. We can measure how fast the amount of compute required is going down.

[Yudkowsky][22:43]

Yeah, that would be a cool thing to measure. It's not obviously a relevant thing to anything important, but it'd be cool to measure.

[Christiano][22:43]

Another way you won't like: we can hold fixed the resources we invest and look at the quality of outputs in any given domain (or even \$ of revenue) and ask how fast it's changing.

[Yudkowsky][22:43]

I wonder what it would say about Go during the age of AlphaGo.

Or what that second metric would say.

[Christiano][22:43]

I think it would be completely fine, and you don't really understand what happened with deep learning in board games. Though I also don't know what happened in much detail, so this is more like a prediction than a retrodiction.

But it's enough of a retrodiction that I shouldn't get too much credit for it.

[Yudkowsky][22:44]

I don't know what result you would consider "completely fine". I didn't have any particular unfine result in mind.

[Christiano][22:45]

oh, sure

if it was just an honest question happy to use it as a concrete case

I would measure the rate of progress in Go by looking at how fast Elo improves with time or increasing R&D spending

[Yudkowsky][22:45]

I mean, I don't have strong predictions about it so it's not yet obviously cruxy to me

[Christiano][22:46]

I'd roughly guess that would continue, and if there were multiple trendlines to extrapolate I'd estimate crossover points based on that

[Yudkowsky][22:47]

suppose this curve is smooth, and we see that sharp Go progress over time happened because Deepmind dumped in a ton of increased R&D spend. you then argue that this cannot happen with AGI because by the time we get there, people will be pushing hard at the frontiers in a competitive environment where everybody's already spending what they can afford, just like in a highly competitive manufacturing industry.

[Christiano][22:47]

the key input to making a prediction for AGZ in particular would be the precise form of the dependence on R&D spending, to try to predict the changes as you shift from a single programmer to a large team at DeepMind, but most reasonable functional forms would be roughly right

Yes, it's definitely a prediction of my view that it's easier to improve things that people haven't spent much money on than things have spent a lot of money on. It's also a separate prediction of my view that people are going to be spending a boatload of money on all of the relevant technologies. Perhaps \$1B/year right now and I'm imagining levels of investment large enough to be essentially bottlenecked on the availability of skilled labor.

[Bensinger][22:48]

(Previous Eliezer-comments about AlphaGo as a break in trend, responding briefly to Miles Brundage: <https://twitter.com/ESRogs/status/1337869362678571008>)

5.7. Legal economic growth

[Yudkowsky][22:49]

Does your prediction change if all hell breaks loose in 2025 instead of 2055?

[Christiano][22:50]

I think my prediction was wrong if all hell breaks loose in 2025, if by "all hell breaks loose" you mean "dyson sphere" and not "things feel crazy"

[Yudkowsky][22:50]

Things feel crazy *in the AI field* and the world ends *less than* 4 years later, well before the world economy doubles.

Why was the Prophecy wrong if the world begins final descent in 2025? The Prophecy requires the world to then last until 2029 while doubling its economic output, after which it is permitted to end, but does not obviously to me forbid the Prophecy to begin coming true in 2025 instead of 2055.

[Christiano][22:52]

yes, I just mean that some important underlying assumptions for the prophecy were violated, I wouldn't put much stock in it at that point, etc.

[Yudkowsky][22:53]

A lot of the issues I have with understanding any of your terminology in concrete Eliezer-language is that it looks to me like the premise-events of your Prophecy are fulfillable in all sorts of ways that don't imply the conclusion-events of the Prophecy.

[Christiano][22:53]

if "things feel crazy" happens 4 years before dyson sphere, then I think we have to be really careful about what crazy means

[Yudkowsky][22:54]

a lot of people looking around nervously and privately wondering if Eliezer was right, while public pravda continues to prohibit wondering anything such thing out loud, so they all go on thinking that they must be wrong.

[Christiano][22:55]

OK, by "things get crazy" I mean like hundreds of billions of dollars of spending at google on automating AI R&D

[Yudkowsky][22:55]

I expect bureaucratic obstacles to prevent much GDP per se from resulting from this.

[Christiano][22:55]

massive scaleups in semiconductor manufacturing, bidding up prices of inputs crazily

[Yudkowsky][22:55]

I suppose that much spending could well increase world GDP by hundreds of billions of dollars per year.

[Christiano][22:56]

massive speculative rises in AI company valuations financing a significant fraction of GWP into AI R&D

(+hardware R&D, +building new clusters, +etc.)

[Yudkowsky][22:56]

like, higher than Tesla? higher than Bitcoin?

both of these things sure did skyrocket in market cap without that having much of an effect on housing stocks and steel production.

[Christiano][22:57]

right now I think hardware R&D is on the order of \$100B/year, AI R&D is more like \$10B/year, I guess I'm betting on something more like trillions? (limited from going higher because of accounting problems and not that much smart money)

I don't think steel production is going up at that point

plausibly going down since you are redirecting manufacturing capacity into making more computers. But probably just staying static while all of the new capacity is going into computers, since cannibalizing existing infrastructure is much more expensive

the original point was: you aren't pulling AlphaZero shit any more, you are competing with an industry that has invested trillions in cumulative R&D

[Yudkowsky][23:00]

is this in hopes of future profit, or because current profits are already in the trillions?

[Christiano][23:01]

largely in hopes of future profit / reinvested AI outputs (that have high market cap), but also revenues are probably in the trillions?

[Yudkowsky][23:02]

this all sure does sound "pretty darn prohibited" on my model, but I'd hope there'd be something earlier than that we could bet on. what does your Prophecy prohibit happening before that sub-prophesied day?

[Christiano][23:02]

To me your model just seems crazy, and you are saying it predicts crazy stuff at the end but no crazy stuff beforehand, so I don't know what's prohibited. Mostly I feel like I'm making positive predictions, of gradually escalating value of AI in lots of different industries

and rapidly increasing investment in AI

I guess your model can be: those things happen, and then one day the AI explodes?

[Yudkowsky][23:03]

the main way you get rapidly increasing investment in AI is if there's some way that AI can produce huge profits without that being effectively bureaucratically prohibited - eg this is where we get huge investments in burning electricity and wasting GPUs on Bitcoin mining.

[Christiano][23:03]

but it seems like you should be predicting e.g. AI quickly jumping to superhuman in lots of domains, and some applications jumping from no value to massive value

I don't understand what you mean by that sentence. Do you think we aren't seeing rapidly increasing investment in AI right now?

or are you talking about increasing investment above some high threshold, or increasing investment at some rate significantly larger than the current rate?

it seems to me like you can pretty seamlessly get up to a few \$100B/year of revenue just by redirecting existing tech R&D

[Yudkowsky][23:05]

so I can imagine scenarios where some version of GPT-5 cloned outside OpenAI is able to talk hundreds of millions of mentally susceptible people into giving away lots of their income, and many regulatory regimes are unable to prohibit this effectively. then AI could be making a profit of trillions and then people would invest corresponding amounts in making new anime waifus trained in erotic hypnosis and findom.

this, to be clear, is not my mainline prediction.

but my sense is that our current economy is mostly not about the 1-day period to design new vaccines, it is about the multi-year period to be allowed to sell the vaccines.

the exceptions to this, like Bitcoin managing to say "fuck off" to the regulators for long enough, are where Bitcoin scales to a trillion dollars and gets massive amounts of electricity and GPU burned on it.

so we can imagine something like this for AI, which earns a trillion dollars, and sparks a trillion-dollar competition.

but my sense is that your model does not work like this.

my sense is that your model is about *general* improvements across the *whole* economy.

[Christiano][23:08]

I think bitcoin is small even compared to current AI...

[Yudkowsky][23:08]

my sense is that we've already built an economy which rejects improvement based on small amounts of cleverness, and only rewards amounts of cleverness large enough to bypass bureaucratic structures. it's not enough to figure out a version of e-gold that's 10% better. e-gold is already illegal. you have to figure out Bitcoin.

what are you going to build? better airplanes? airplane costs are mainly regulatory costs. better medtech? mainly regulatory costs. better houses? building houses is illegal anyways.

where is the room for the general AI revolution, short of the AI being literally revolutionary enough to overthrow governments?

[Christiano][23:10]

factories, solar panels, robots, semiconductors, mining equipment, power lines, and "factories" just happens to be one word for a thousand different things

I think it's reasonable to think some jurisdictions won't be willing to build things but it's kind of improbable as a prediction for the whole world. That's a possible source of shorter-term predictions?

also computers and the 100 other things that go in datacenters

[Yudkowsky][23:12]

The whole developed world rejects open borders. The regulatory regimes all make the same mistakes with an almost perfect precision, the kind of coordination that human beings could never dream of when trying to coordinate on purpose.

if the world lasts until 2035, I could perhaps see deepnets becoming as ubiquitous as computers were in... 1995? 2005? would that fulfill the terms of the Prophecy? I think it doesn't; I think your Prophecy requires that early AGI tech be that ubiquitous so that AGI tech will have trillions invested in it.

[Christiano][23:13]

what is AGI tech?

the point is that there aren't important drivers that you can easily improve a lot

[Yudkowsky][23:14]

for purposes of the Prophecy, AGI tech is that which, scaled far enough, ends the world; this must have trillions invested in it, so that the trajectory up to it cannot look like pulling an AlphaGo. no?

[Christiano][23:14]

so it's relevant if you are imagining some piece of the technology which is helpful for general problem solving or something but somehow not helpful for all of the things people are doing with ML, to me that seems unlikely since it's all the same stuff

surely AGI tech should at least include the use of AI to automate AI R&D

regardless of what you arbitrarily decree as "ends the world if scaled up"

[Yudkowsky][23:15]

only if that's the path that leads to destroying the world?

if it isn't on that path, who cares Prophecy-wise?

[Christiano][23:15]

also I want to emphasize that "pull an AlphaGo" is what happens when you move from SOTA being set by an individual programmer to a large lab, you don't need to be investing trillions to avoid that

and that the jump is still more like a few years

but the prophecy does involve trillions, and my view gets more like your view if people are jumping from \$100B of R&D ever to \$1T in a single year

5.8. TPUs and GPUs, and automating AI R&D

[Yudkowsky][23:17]

I'm also wondering a little why the emphasis on "trillions". it seems to me that the terms of your Prophecy should be fulfillable by AGI tech being merely as ubiquitous as modern computers, so that many competing companies invest mere hundreds of billions in the equivalent of hardware plants. it is legitimately hard to get a chip with 50% better transistors ahead of TSMC.

[Christiano][23:17]

yes, if you are investing hundreds of billions then it is hard to pull ahead (though could still happen)

(since the upside is so much larger here, no one cares that much about getting ahead of TSMC since the payoff is tiny in the scheme of the amounts we are discussing)

[Yudkowsky][23:18]

which, like, doesn't prevent Google from tossing out TPUs that are pretty significant jumps on GPUs, and if there's a specialized application of AGI-ish tech that is especially key, you can have everything behave smoothly and still get a jump that way.

[Christiano][23:18]

I think TPUs are basically the same as GPUs

probably a bit worse

(but GPUs are sold at a 10x markup since that's the size of nvidia's lead)

[Yudkowsky][23:19]

noted; I'm not enough of an expert to directly contradict that statement about TPUs from my own knowledge.

[Christiano][23:19]

(though I think TPUs are nevertheless leased at a slightly higher price than GPUs)

[Yudkowsky][23:19]

how does Nvidia maintain that lead and 10x markup? that sounds like a pretty un-Paul-ish state of affairs given Bitcoin prices never mind AI investments.

[Christiano][23:20]

nvidia's lead isn't worth that much because historically they didn't sell many gpus

(especially for non-gaming applications)

their R&D investment is relatively large compared to the \$ on the table

my guess is that their lead doesn't stick, as evidenced by e.g. Google very quickly catching up

[Yudkowsky][23:21]

parenthetically, does this mean - and I don't necessarily predict otherwise - that you predict a drop in Nvidia's stock and a drop in GPU prices in the next couple of years?

[Christiano][23:21]

nvidia's stock may do OK from riding general AI boom, but I do predict a relative fall in nvidia compared to other AI-exposed companies

(though I also predicted google to more aggressively try to compete with nvidia for the ML market and think I was just wrong about that, though I don't really know any details of the area)

I do expect the cost of compute to fall over the coming years as nvidia's markup gets eroded to be partially offset by increases in the cost of the underlying silicon (though that's still bad news for nvidia)

[Yudkowsky][23:23]

I parenthetically note that I think the Wise Reader should be justly impressed by predictions that come true about relative stock price changes, even if Eliezer has not explicitly contradicted those predictions before they come true. there are bets you can win without my having to bet against you.

[Christiano][23:23]

you are welcome to counterpredict, but no saying in retrospect that reality proved you right if you don't 😊

otherwise it's just me vs the market

[Yudkowsky][23:24]

I don't feel like I have a counterprediction here, but I think the Wise Reader should be impressed if you win vs. the market.

however, this does require you to name in advance a few "other AI-exposed companies".

[Christiano][23:25]

Note that I made the same bet over the last year---I make a large AI bet but mostly moved my nvidia allocation to semiconductor companies. The semiconductor part of the portfolio is up 50% while nvidia is up 70%, so I lost that one. But that just means I like the bet even more next year.

happy to use nvidia vs tsmc

[Yudkowsky][23:25]

there's a lot of noise in a 2-stock prediction.

[Christiano][23:25]

I mean, it's a 1-stock prediction about nvidia

[Yudkowsky][23:26]

but your funeral or triumphal!

[Christiano][23:26]

indeed 😊

anyway

I expect all of the \$ amounts to be much bigger in the future

[Yudkowsky][23:26]

yeah, but using just TSMC for the opposition exposes you to I dunno Chinese invasion of Taiwan

[Christiano][23:26]

yes

also TSMC is not that AI-exposed

I think the main prediction is: eventual move away from GPUs, nvidia can't maintain that markup

[Yudkowsky][23:27]

"Nvidia can't maintain that markup" sounds testable, but is less of a win against the market than predicting a relative stock price shift. (Over what timespan? Just the next year sounds quite fast for that kind of prediction.)

[Christiano][23:27]

regarding your original claim: if you think that it's plausible that AI will be doing all of the AI R&D, and that will be accelerating continuously from 12, 6, 3 month "doubling times," but that we'll see a discontinuous change in the "path to doom," then that would be harder to generate predictions about

yes, it's hard to translate most predictions about the world into predictions about the stock market

[Yudkowsky][23:28]

this again sounds like it's not written in Eliezer-language.

what does it mean for "AI will be doing all of the AI R&D"? that sounds to me like something that happens after the end of the world, hence doesn't happen.

[Christiano][23:29]

that's good, that's what I thought

[Yudkowsky][23:29]

I don't necessarily want to sound very definite about that in advance of understanding what it *means*

[Christiano][23:29]

I'm saying that I think AI will be automating AI R&D gradually, before the end of the world
yeah, I agree that if you reject the construct of "how fast the AI community makes progress"
then it's hard to talk about what it means to automate "progress"
and that may be hard to make headway on
though for cases like AlphaGo (which started that whole digression) it seems easy enough to
talk about elo gain per year
maybe the hard part is aggregating across tasks into a measure you actually care about?

[Yudkowsky][23:30]

up to a point, but yeah. (like, if we're taking Elo high above human levels and restricting our
measurements to a very small range of frontier AIs, I quietly wonder if the measurement is
still measuring quite the same thing with quite the same robustness.)

[Christiano][23:31]

I agree that elo measurement is extremely problematic in that regime

5.9. Smooth exponentials vs. jumps in income

[Yudkowsky][23:31]

so in your worldview there's this big emphasis on things that must have been deployed and
adopted widely to the point of already having huge impacts

and in my worldview there's nothing very surprising about people with a weird powerful
prototype that wasn't used to automate huge sections of AI R&D because the previous
versions of the tech weren't useful for that or bigcorps didn't adopt it.

[Christiano][23:32]

I mean, Google is already 1% of the US economy and in this scenario it and its peers are
more like 10-20%? So wide adoption doesn't have to mean that many people. Though I also
do predict much wider adoption than you so happy to go there if it's happy for predictions.

I don't really buy the "weird powerful prototype"

[Yudkowsky][23:33]

yes. I noticed.

you would seem, indeed, to be offering large quantities of it for short sale.

[Christiano][23:33]

and it feels like the thing you are talking about ought to have some precedent of some kind,
of weird powerful prototypes that jump straight from "does nothing" to "does something
impactful"

like if I predict that AI will be useful in a bunch of domains, and will get there by small steps, you should either predict that won't happen, or else also predict that there will be some domains with weird prototypes jumping to giant impact?

[Yudkowsky][23:34]

like an electrical device that goes from "not working at all" to "actually working" as soon as you screw in the attachments for the electrical plug.

[Christiano][23:34]

(clearly takes more work to operationalize)

I'm not sure I understand that sentence, hopefully it's clear enough why I expect those discontinuities?

[Yudkowsky][23:34]

though, no, that's a facile bad analogy.

a better analogy would be an AI system that only starts working after somebody tells you about batch normalization or LAMB learning rate or whatever.

[Christiano][23:36]

sure, which I think will happen all the time for individual AI projects but not for sota because the projects at sota have picked the low hanging fruit, it's not easy to get giant wins

[Yudkowsky][23:36]

like if I predict that AI will be useful in a bunch of domains, and will get there by small steps, you should either predict that won't happen, or else also predict that there will be some domains with weird prototypes jumping to giant impact?

in the latter case, has this Eliezer-Prophecy already had its terms fulfilled by AlphaFold 2, or do you say nay because AlphaFold 2 hasn't doubled GDP?

[Christiano][23:37]

(you can also get giant wins by a new competitor coming up at a faster rate of progress, and then we have more dependence on whether people do it when it's a big leap forward or slightly worse than the predecessor, and I'm betting on the latter)

I have no idea what AlphaFold 2 is good for, or the size of the community working on it, my guess would be that its value is pretty small

we can try to quantify

like, I get surprised when \$X of R&D gets you something whose value is much larger than \$X

I'm not surprised at all if \$X of R&D gets you <<\$X, or even like 10*\$X in a given case that was selected for working well

hopefully it's clear enough why that's the kind of thing a naive person would predict

[Yudkowsky][23:38]

so a thing which Eliezer's Prophecy does not mandate per se, but sure does permit, and is on the mainline especially for nearer timelines, is that the world-ending prototype had no prior prototype containing 90% of the technology which earned a trillion dollars.

a lot of Paul's Prophecy seems to be about forbidding this.

is that a fair way to describe your own Prophecy?

[Christiano][23:39]

I don't have a strong view about "containing 90% of the technology"

the main view is that whatever the "world ending prototype" does, there were earlier systems that could do practically the same thing

if the world ending prototype does something that lets you go foom in a day, there was a system years earlier that could foom in a month, so that would have been the one to foom

[Yudkowsky][23:41]

but, like, the world-ending thing, according to the Prophecy, must be squarely in the middle of a class of technologies which are in the midst of earning trillions of dollars and having trillions of dollars invested in them. it's not enough for the Worldender to be definitionally somewhere in that class, because then it could be on a weird outskirt of the class, and somebody could invest a billion dollars in that weird outskirt before anybody else had invested a hundred million, which is forbidden by the Prophecy. so the Worldender has got to be right in the middle, a plain and obvious example of the tech that's already earning trillions of dollars. ...y/n?

[Christiano][23:42]

I agree with that as a prediction for some operationalization of "a plain and obvious example," but I think we could make it more precise / it doesn't feel like it depends on the fuzziness of that

I think that if the world can end out of nowhere like that, you should also be getting \$100B/year products out of nowhere like that, but I guess you think not because of bureaucracy

like, to me it seems like our views stake out predictions about codex, where I'm predicting its value will be modest relative to R&D, and the value will basically improve from there with a nice experience curve, maybe something like ramping up quickly to some starting point <\$10M/year and then doubling every year thereafter, whereas I feel like you are saying more like "who knows, could be anything" and so should be surprised each time the boring thing happens

[Yudkowsky][23:45]

the concrete example I give is that the World-Ending Company will be able to use the same tech to build a true self-driving car, which would in the natural course of things be approved for sale a few years later after the world had ended.

[Christiano][23:46]

but self-driving cars seem very likely to already be broadly deployed, and so the relevant question is really whether their technical improvements can also be deployed to those cars?

(or else maybe that's another prediction we disagree about)

[Yudkowsky][23:47]

I feel like I would indeed not have the right to feel very surprised if Codex technology stagnated for the next 5 years, nor if it took a massive leap in 2 years and got ubiquitously adopted by lots of programmers.

yes, I think that's a general timeline difference there

re: self-driving cars

I might be talkable into a bet where you took "Codex tech will develop like *this*" and I took the side "literally anything else but that"

[Christiano][23:48]

I think it would have to be over/under, I doubt I'm more surprised than you by something failing to be economically valuable, I'm surprised by big jumps in value

seems like it will be tough to work

[Yudkowsky][23:49]

well, if I was betting on something taking a big jump in income, I sure would bet on something in a relatively unregulated industry like Codex or anime waifus.

but that's assuming I made the bet at all, which is a hard sell when the bet is about the Future, which is notoriously hard to predict.

[Christiano][23:50]

I guess my strongest take is: if you want to pull the thing where you say that future developments proved you right and took unreasonable people like me by surprise, you've got to be able to say *something* in advance about what you expect to happen

[Yudkowsky][23:51]

so what if neither of us are surprised if Codex stagnates for 5 years, you win if Codex shows a smooth exponential in income, and I win if the income looks... jumpier? how would we quantify that?

[Christiano][23:52]

Codex also does seem a bit unfair to you in that it may have to be adopted by lots of programmers which could slow things down a lot even if capabilities are pretty jumpy

(though I think in fact usefulness and not merely profit will basically just go up smoothly, with step sizes determined by arbitrary decisions about when to release something)

[Yudkowsky][23:53]

I'd also be concerned about unfairness to me in that earnable income is not the same as the gains from trade. If there's more than 1 competitor in the industry, their earnings from Codex may be much less than the value produced, and this may not change much with improvements in the tech.

5.10. Late-stage predictions

[Christiano][23:53]

I think my main update from this conversation is that you don't really predict someone to come out of nowhere with a model that can earn a lot of \$, even if they could come out of nowhere with a model that could end the world, because of regulatory bottlenecks and nimbyism and general sluggishness and unwillingness to do things
does that seem right?

[Yudkowsky][23:55]

Well, and also because the World-ender is "the first thing that scaled with compute" and/or "the first thing that ate the real core of generality" and/or "the first thing that went over neutron multiplication factor 1".

[Christiano][23:55]

and so that cuts out a lot of the easily-specified empirical divergences, since "worth a lot of \$" was the only general way to assess "big deal that people care about" and avoiding disputes like "but Zen was mostly developed by a single programmer, it's not like intense competition"

yeah, that's the real disagreement it seems like we'd want to talk about

but it just doesn't seem to lead to many prediction differences in advance?

I totally don't buy any of those models, I think they are bonkers

would love to bet on that

[Yudkowsky][23:56]

Prolly but I think the from-my-perspective-weird talk about GDP is probably concealing *some* kind of important crux, because caring about GDP still feels pretty alien to me.

[Christiano][23:56]

I feel like getting up to massive economic impacts without seeing "the real core of generality" seems like it should also be surprising on your view

like if it's 10 years from now and AI is a pretty big deal but no crazy AGI, isn't that surprising?

[Yudkowsky][23:57]

Mildly but not too surprising, I would imagine that people had built a bunch of neat stuff with gradient descent in realms where you could get a long way on self-play or massively collectible datasets.

[Christiano][23:58]

I'm fine with the crux being something that doesn't lead to any empirical disagreements, but in that case I just don't think you should claim credit for the worldview making great predictions.

(or the countervailing worldview making bad predictions)

[Yudkowsky][23:59]

stuff that we could see then: self-driving cars (10 years is enough for regulatory approval in many countries), super Codex, GPT-6 powered anime waifus being an increasingly loud source of (arguably justified) moral panic and a hundred-billion-dollar industry

[Christiano][23:59]

another option is "10% GDP GWP growth in a year, before doom"

I think that's very likely, though might be too late to be helpful

[Yudkowsky][0:01] (next day, Sep. 15)

see, that seems genuinely hard unless somebody gets GPT-4 far head of any political opposition - I guess all the competent AGI groups lean solidly liberal at the moment? - and uses it to fake massive highly-persuasive sentiment on Twitter for housing liberalization.

[Christiano][0:01] (next day, Sep. 15)

so seems like a bet?

but you don't get to win until doom 😞

[Yudkowsky][0:02] (next day, Sep. 15)

I mean, as written, I'd want to avoid cases like 10% growth on paper while recovering from a pandemic that produced 0% growth the previous year.

[Christiano][0:02] (next day, Sep. 15)

yeah

[Yudkowsky][0:04] (next day, Sep. 15)

I'd want to check the current rate (5% iirc) and what the variance on it was, 10% is a little low for surety (though my sense is that it's a pretty darn smooth graph that's hard to perturb)

if we got 10% in a way that was clearly about AI tech becoming that ubiquitous, I'd feel relatively good about nodding along and saying, "Yes, that is like unto the beginning of Paul's Prophecy" not least because the timelines had been that long at all.

[Christiano][0:05] (next day, Sep. 15)

like 3-4%/year right now

random wikipedia number is 5.5% in 2006-2007, 3-4% since 2010

4% 1995-2000

[Yudkowsky][0:06] (next day, Sep. 15)

I don't want to sound obstinate here. My model does not *forbid* that we dwiddle around on the AGI side while gradient descent tech gets its fingers into enough separate weakly-generalizing pies to produce 10% GDP growth, but I'm happy to say that this sounds much more like Paul's Prophecy is coming true.

[Christiano][0:07] (next day, Sep. 15)

ok, we should formalize at some point, but also need the procedure for you getting credit given that it can't resolve in your favor until the end of days

[Yudkowsky][0:07] (next day, Sep. 15)

Is there something that sounds to you like Eliezer's Prophecy which we can observe before the end of the world?

[Christiano][0:07] (next day, Sep. 15)

when you will already have all the epistemic credit you need

not on the "simple core of generality" stuff since that apparently immediately implies end of world

maybe something about ML running into obstacles en route to human level performance?

or about some other kind of discontinuous jump even in a case where people care, though there seem to be a few reasons you don't expect many of those

[Yudkowsky][0:08] (next day, Sep. 15)

depends on how you define "immediately"? it's not *long* before the end of the world, but in some sad scenarios there is some tiny utility to you declaring me right 6 months before the end.

[Christiano][0:09] (next day, Sep. 15)

I care a lot about the 6 months before the end personally

though I do think probably everything is more clear by then independent of any bet; but I guess you are more pessimistic about that

[Yudkowsky][0:09] (next day, Sep. 15)

I'm not quite sure what I'd do in them, but I may have worked something out before then, so I care significantly in expectation if not in particular.

I am more pessimistic about other people's ability to notice what reality is screaming in their faces, yes.

[Christiano][0:10] (next day, Sep. 15)

if we were to look at various scaling curves, e.g. of loss vs model size or something, do you expect those to look distinctive as you hit the "real core of generality"?

[Yudkowsky][0:10] (next day, Sep. 15)

let me turn that around: if we add transformers into those graphs, do they jump around in a way you'd find interesting?

[Christiano][0:11] (next day, Sep. 15)

not really

[Yudkowsky][0:11] (next day, Sep. 15)

is that because the empirical graphs don't jump, or because you don't think the jumps say much?

[Christiano][0:11] (next day, Sep. 15)

but not many good graphs to look at (I just have one in mind), so that's partly a prediction about what the exercise would show

I don't think the graphs jump much, and also transformers come before people start evaluating on tasks where they help a lot

[Yudkowsky][0:12] (next day, Sep. 15)

It would not terribly contradict the terms of my Prophecy if the World-ending tech began by not producing a big jump on existing tasks, but generalizing to some currently not-so-popular tasks where it scaled much faster.

[Christiano][0:13] (next day, Sep. 15)

eh, they help significantly on contemporary tasks, but it's just not a huge jump relative to continuing to scale up model sizes

or other ongoing improvements in architecture

anyway, should try to figure out something, and good not to finalize a bet until you have some way to at least come out ahead, but I should sleep now

[Yudkowsky][0:14] (next day, Sep. 15)

yeah, same.

Thing I want to note out loud lest I forget ere I sleep: I think the real world is full of tons and tons of technologies being developed as unprecedented prototypes in the midst of big fields, because the key thing to invest in wasn't the competitively explored center. Wright Flyer vs all expenditures on Traveling Machine R&D. First atomic pile and bomb vs all Military R&D.

This is one reason why Paul's Prophecy seems fragile to me. You could have the preliminaries come true as far as there being a trillion bucks in what looks like AI R&D, and then the WorldEnder is a weird prototype off to one side of that. saying "But what about the rest of that AI R&D?" is no more a devastating retort to reality than looking at AlphaGo and saying "But weren't other companies investing billions in Better Software?" Yeah but it was a big playing field with lots of different kinds of Better Software and no other medium-sized team of 15 people with corporate TPU backing was trying to build a system just like AlphaGo, even though multiple small outfits were trying to build prestige-earning gameplayers. Tech advancements very very often occur in places where investment wasn't dense enough to guarantee overlap.

6. Follow-ups on "Takeoff Speeds"

6.1. Eliezer Yudkowsky's commentary

[Yudkowsky][17:25] (Sep. 15)

Further comment that occurred to me on "takeoff speeds" if I've better understood the main thesis now: its hypotheses seem to include a perfectly anti-Thielian setup for AGI.

Thiel has a running thesis about how part of the story behind the Great Stagnation and the decline in innovation that's about atoms rather than bits - the story behind "we were promised flying cars and got 140 characters", to cite the classic Thielian quote - is that people stopped believing in "[secrets](#)".

Thiel suggests that you have to believe there are knowable things that aren't yet widely known - not just things that everybody already knows, plus mysteries that nobody will ever know - in order to be motivated to go out and innovate. Culture in developed countries shifted to label this kind of thinking rude - or rather, even ruder, even less tolerated than it had been decades before - so innovation decreased as a result.

The central hypothesis of "takeoff speeds" is that at the time of serious AGI being developed, it is perfectly anti-Thielian in that it is devoid of secrets in that sense. It is not permissible (on this viewpoint) for it to be the case that there is a lot of AI investment into AI that is directed not quite at the key path leading to AGI, such that somebody could spend \$1B on compute for the key path leading to AGI before anybody else had spent \$100M on that. There cannot exist any secret like that. The path to AGI will be known; everyone, or a wide variety of powerful actors, will know how profitable that path will be; the surrounding industry will be capable of acting on this knowledge, and will have actually been acting on it as early as possible; multiple actors are already investing in every tech path that would in fact be profitable (and is known to any human being at all), as soon as that R&D opportunity becomes available.

And I'm not saying this is an inconsistent world to describe! I've written science fiction set in this world. I called it "[dath ilan](#)". It's a hypothetical world that is actually full of smart people in economic equilibrium. If anything like Covid-19 appears, for example, the governments and public-good philanthropists there have already set up prediction markets (which are not illegal, needless to say); and of course there are mRNA vaccine factories already built and ready to go, because somebody already calculated the profits from fast vaccines would be very high in case of a pandemic (no artificial price ceilings in this world, of course); so as soon as the prediction markets started calling the coming pandemic conditional on no vaccine, the mRNA vaccine factories were already spinning up.

This world, however, is not Earth.

On Earth, major chunks of technological progress quite often occur *outside* of a social context where everyone knew and agreed in advance on which designs would yield how much expected profit and many overlapping actors competed to invest in the most actually-promising paths simultaneously.

And that is why you can read [Inadequate Equilibria](#), and then read this essay on takeoff speeds, and go, "Oh, yes, I recognize this; it's written inside the Modesty worldview; in particular, the imagination of an adequate world in which there is a perfect absence of Thielian secrets or unshared knowable knowledge about fruitful development pathways. This is the same world that already had mRNA vaccines ready to spin up on day one of the Covid-19 pandemic, because markets had correctly forecasted their option value and investors had acted on that forecast unimpeded. Sure would be an interesting place to live! But we don't live there."

Could we perhaps end up in a world where the path to AGI is in fact not a Thielian secret, because in fact the first accessible path to AGI happens to lie along a tech pathway that already delivered large profits to previous investors who summed a lot of small innovations, a la experience with chipmaking, such that there were no large innovations just lots and lots of small innovations that yield 10% improvement annually on various tech benchmarks?

I think that even in this case we will get weird, discontinuous, and fatal behaviors, and I could maybe talk about that when discussion resumes. But it is not ruled out to me that the first accessible pathway to AGI could happen to lie in the further direction of some road that was already well-traveled, already yielded much profit to now-famous tycoons back when its

first steps were Thielian secrets, and hence is now replete with dozens of competing chasers for the gold rush.

It's even imaginable to me, though a bit less so, that the first path traversed to real actual pivotal/powerful/lethal AGI, happens to lie literally actually squarely in the central direction of the gold rush. It sounds a little less like the tech history I know, which is usually about how someone needed to swerve a bit and the popular gold-rush forecasts weren't quite right, but maybe that is just a selective focus of history on the more interesting cases.

Though I remark that - even supposing that getting to big AGI is literally as straightforward and yet as difficult as falling down a semiconductor manufacturing roadmap (as otherwise the biggest actor to first see the obvious direction could just rush down the whole road) - well, TSMC does have a bit of an unshared advantage right now, if I recall correctly. And Intel had a bit of an advantage before that. So that happens even when there's competitors competing to invest billions.

But we can imagine that doesn't happen either, because instead of needing to build a whole huge manufacturing plant, there's just lots and lots of little innovations adding up to every key AGI threshold, which lots of actors are investing \$10 million in at a time, and everybody knows which direction to move in to get to more serious AGI and they're right in this shared forecast.

I am willing to entertain discussing this world and the sequelae there - I do think everybody still dies in this case - but I would not have this particular premise thrust upon us as a default, through a not-explicitly-spoken pressure against being so immodest and inequitable as to suppose that any Thielian knowable-secret will exist, or that anybody in the future gets as far ahead of others as today's TSMC or today's Deepmind.

We are, in imagining this world, imagining a world in which AI research has become drastically unlike today's AI research in a direction drastically different from the history of many other technologies.

It's not literally unprecedented, but it's also not a default environment for big moments in tech progress; it's narrowly preceded for *particular* industries with high competition and steady benchmark progress driven by huge investments into a sum of many tiny innovations.

So I can entertain the scenario. But if you want to claim that the social situation around AGI *will* drastically change in this way you foresee - not just that it *could* change in that direction, if somebody makes a big splash that causes everyone else to reevaluate their previous opinions and arrive at yours, but that this social change *will* occur and you know this now - and that the prerequisite tech path to AGI is known to you, and forces an investment situation that looks like the semiconductor industry - then your "What do you think you know and how do you think you know it?" has some significant explaining to do.

Of course, I do appreciate that such a thing could be knowable, and yet not known to me. I'm not so silly as to disbelieve in secrets like that. They're all over the actual history of technological progress on our actual Earth.

Almost everyone should be less afraid of lawsuits

One sad feature of modern American society is that many people, especially those tied to big institutions, don't help each other out because of a fear of lawsuits. Employers don't give meaningful [references](#), or ever tell their [rejected interviewees](#) how they could improve their skills. Abuse victims keep silent, in case someone on their abuser's side files a defamation case. Doctors prescribe unnecessary, expensive tests as "[defensive medicine](#)". Inventions don't get built, in case there's a patent lawsuit. I'm not an attorney myself, but my best guess is that letting litigation fears stop you is often a mistake, and I've given this advice to friends several times before. Here's why:

Almost all lawsuit threats never happen

Threats are easy - anyone can threaten to sue anyone else, with two minutes of time and a smartphone. Actually suing is much harder. Outside of [small claims court](#), hiring an attorney will usually cost tens of thousands of dollars, at least. Litigating a case takes months or even years, while angry feelings often go away after a few weeks. The person suing will have to give up a lot. Instead of playing games or taking a vacation or putting in extra hours at work, they will have to do legal research and give testimony. Most people are distracted by their families, their career, their hobbies, and their lives, and will (often rationally) eventually give up, rather than remaining obsessed with whatever the case was about.

Lawsuit mitigation can be expensive

Doctors as a profession are traditionally concerned with legal risk. But the total value of medical malpractice claims is around [\\$5 billion per year](#) in the US - compared to healthcare spending of [\\$3,800 billion](#), malpractice lawyer fees of [\\$3 billion](#), and "defensive medicine" costs of [\\$45 billion](#) according to one study. Likewise, the cost to media companies and journalists of not publishing articles to stop defamation suits surely exceeds that of the [few dozen defamation cases](#) against them every year (stats from the UK, but British defamation law is widely considered plaintiff-friendly). The cost of things like [rape victims staying silent](#) because of defamation threats is hard to quantify, but clearly also bad. Some of these costs will fall on you directly, but often many more are [externalized](#); eg., the cost to a patient of not giving them the right treatment out of fear can be enormous.

Many legal risks are hypothetical or imaginary

Since most people aren't legal specialists, the "knowledge" that something is a legal risk can be passed around from place to place without ever being checked, like [many urban legends](#) or [droplet disease transmission](#) or [stories about the amount of iron in spinach](#). For example, a lot of people "hear" that IQ tests are illegal in hiring. But not only is this [not true](#), any company can easily buy one right now - "the [CCAT](#)" is a pre-employment aptitude test that measures an individual's aptitude, or ability to solve problems, digest and apply information, learn new skills, and think critically. Individuals with high aptitude are more likely to be quick learners and high performers than are individuals with low aptitude." Likewise, it's easy to make up a story where

something might be hypothetically risky, even if it's never happened in court, and hard to do enough research to make sure it isn't.

Most lawsuit types are rare

One study found around [750,000](#) civil lawsuits per year in big counties, with a population around 75 million, for a total annual risk of ~1% per person. But the distribution of case types is [thin-tailed](#). Courts and lawyers are somewhat machine-like, and mostly handle the same mundane problems over and over again (this is also true for [criminal cases](#)). A majority of all the cases were either about car crashes or debt collection. Another ~9% were mortgage foreclosures. The incidence of all employment lawsuits was 1 per ~10,000 person-years; fraud lawsuits were 1 per ~5,000; product liability suits were 1 per ~6,000. The number of cases involving something like the [GPL](#) license is tiny, despite its ubiquitous use in software. Most lawyers have little quantitative training, and won't try to crunch the numbers to calculate probabilities or expected values.

You might get sued anyway

If someone is rich, idle, and vindictive enough to really sue you, your paranoid lawsuit-proofing actions might not help. They can often sue even without a good case, and the cost to you might not be that different; time costs and lawyer fees are often similar whether you win or lose. In the infamous [Prenda Law](#) case, a team of unscrupulous lawyers took advantage of this to put porn on torrent sites, wait for it to be pirated, send mass lawsuit threats asking to be paid off (where the price was a lot, but less than that of hiring an attorney), and then eventually drop any case that someone contested as a form of extortion. Groups will sometimes use laws like [CEQA](#) as a form of legal blackmail - their goal is not to correct the "problem" they cite, but to get you to go away or pay them off. CEQA is so complex, covering over a hundred "environmental" topics (where the "environmental damage" has often been something like not providing enough parking spots), that a determined litigant can almost always find *something* to object to no matter what one does.

Lawyers aren't your boss

"Lawyer" is traditionally a high-status job, and most people look to lawyers with some degree of deference and authority. When a lawyer says "don't do X, you might get sued", people usually listen. But lawyers are institutionally trained and personally incentivized to be conservative, largely because of [asymmetric justice](#) - if you get sued and the lawyer didn't speak up, they could be blamed, while lawyers are almost never blamed for shooting down an otherwise high-value idea out of inaccurate risk assessment. Just like [airports and missing a flight](#), an organization that always listens to lawyers is being more cautious than the optimum.

Just dragging it out often works

Donald Trump, despite being a cackling cartoon villain, was able to get as far as he did largely because he wasn't afraid of risk, including legal risk. He often did things that were pretty clearly illegal, he and his companies were [sued all the time](#), and he mostly just kept fighting until the case was dropped or otherwise shrugged it off. People who dislike Trump, who wonder how anyone like that could get elected President, should

take notice that this tactic *worked*, that even someone like Donald Trump wasn't stopped by litigation despite being sued by everybody at a million miles an hour. In business, for example, if a competitor sues you, by the time the suit is resolved your company will likely have succeeded or failed anyway.

The world is kind of in a mess right now

If your life is completely perfect, then you should try not to ever change anything, since by assumption any change can only make something worse. On the flip side, if your life is already as bad as it could possibly be, then any change must logically be good. Just in case anyone has been living under a rock, things in general are [going wrong right now](#), and it makes sense to take risks (increase variance) if on your current path you lose by default. On a smaller scale, many individuals and companies are in a [bad situation](#), where they have big problems with no obvious solution or are trending towards bankruptcy, and could stand to benefit from more risk-taking.

The worst case isn't that awful

If you are sued, try everything you can, lose anyway, and can't afford to pay the judgment, your wages can be garnished, but only up to [25%](#) (in some states less, in Texas [nothing at all](#)). That's bad, but it's not infinitely bad; it roughly works out to [-0.1 standard deviations](#) on typical happiness/well-being metrics, and is comparable to the cost of routine things like divorce (without the romantic heartbreak!) or changing careers. For a business or other organization, the worst case is usually bankruptcy, but businesses go bankrupt all the time when markets change or key employees leave. People are vastly more afraid of lawsuits than other things with similar outcomes, like taking on six figures of student debt to [go to law school](#) without having researched the job market. Please don't do that. :)

The Rationalists of the 1950s (and before) also called themselves “Rationalists”

TLDR

- There's an organization based in London called the [Rationalist Association](#). It was founded in 1885. Historically, it focused on publishing books and articles related to atheism and science, including works by Darwin, Bertrand Russell, J. B. S. Haldane, George Bernard Shaw, H. G. Wells, and Karl Popper.
- The topics covered overlap with the present-day [rationalist movement](#) (centered on Lesswrong). They include religion and atheism, philosophy (especially philosophy of science and ethics), evolution, and psychology.
- According to [Wikipedia](#), membership of the Rationalist Association peaked in 1959 with more than 5000 members and with Bertrand Russell as President.
- This post displays some covers of Rationalist Association publications, and links to full-text articles and other resources.
- Prior to reading this [biography](#), I hadn't heard of these earlier rationalists. So I did some quick and shallow research. I'd be curious to know more about this history. It might be worth adding a note to this LW Wiki [entry](#).

Past covers of the Rationalist Association publications

1896 Cover

- This is a journal for short articles called the “Agnostic Annual” (later “Rationalist Annual”). The full text is [here](#).
- Some quotes from the article “Mind as controlled by matter” below.

1896

THE
**AGNOSTIC
ANNUAL**

— 1896. —

EDITED BY CHARLES A. WATTS.

Balfour's "Foundations of Belief": An Agnostic Rejoinder

S. LAING

Agnosticism and its Equivalents AMOS WATERS

The Man, Christ Jesus: The Germ of the Christian Myth

J. ALLANSON PICTON

Psyche: A Poem W. STEWART ROSS (SALADIN)

Mind as Controlled by Matter CONSTANCE E. PLUMPTRE

The Faiths of Our Forefathers CHARLES WATTS

An Agnostic View of Theism and Monism. R. BITHELL, B.Sc., PH.D.

The Old Testament Library F. J. GOULD

Immortality W. A. LEONARD

The Physiological Bias of Religious Leaders

FURNEAUX JORDAN, F.R.C.S.

1938 Cover

Includes articles by Bertrand Russell and evolutionary biologist and [proto-transhumanist](#) J. B. S. Haldane.

One Shilling Net

The

RATIONALIST ANNUAL

1938

BERTRAND RUSSELL

MY RELIGIOUS REMINISCENCES

PROF. J. B. S. HALDANE

BEYOND EINSTEIN

SIR P. CHALMERS MITCHELL

THE DECAY OF BELIEF IN GOD

LLEWELYN POWYS

THE SUPERSTITIOUS SOPHISTICATIONS
OF CHRISTIANITY

ERNEST THURTE, M.P.

A SECULAR SAINT: NOTES ON
LAWRENCE OF ARABIA

WALTER TROUGHTON

HERBERT SPENCER'S LAST YEARS:
SOME PERSONAL RECOLLECTIONS

DR. HORACE J. BRIDGES

RATIONALISM IN AMERICA

GERALD BULLETT

REASON AND RELIGION

S. K. RATCLIFFE

WHAT THEY BELIEVE

ARCHIBALD ROBERTSON

IS THERE A HALF-WAY HOUSE?

JOHN LANGDON-DAVIES

THE FUTURE OF RELIGION
IN SPAIN

J. W. POYNTER

THE DISASTROUS PONTIFICATE
OF PIUS XI

A. GOWANS WHYTE

"HAVING PASSED AWAY,
WILL NEVER RETURN"

JOHN ROWLAND

HERESY IN MODERN FICTION

1954 Cover

THE
**Rationalist
Annual**

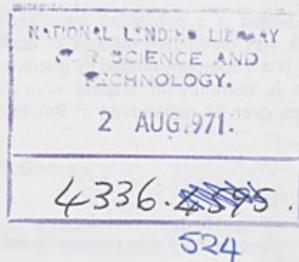
- BERTRAND RUSSELL
Are the World's Troubles Due to Decay of Faith ?
- J. B. S. HALDANE A Rationalist with a Halo
- T. H. PEAR
The Place of the Psychologist in the Community
- SOMERSET MAUGHAM The Judgment Seat
- PATRICK NOWELL-SMITH
Psycho-analysis and Moral Language
- W. E. SWINTON Missing Links
- DONALD G. MACRAE
Existential Philosophy and M. Sartre
- F. H. GEORGE
"Public" and "Private" Knowledge
- HECTOR HAWTON The Logic of Total War

I 954

- This is for a magazine called The Humanist, published by the Rationalist Association. It has since published articles by Richard Dawkins and Daniel Dennett.
- Contents: “Rationalist Tasks”, “Motoring Safety”, “Quasi-rational Quarrelling” and the first appearance of the famous Philip Larkin [poem](#) that begins “They fuck you up, your mum and dad”.

August 1971 Ten New Pence

86
8



HUMANIST HUMANIST

RATIONALIST TASKS
BARBARA WOOTTON

HUMANIST HUMANIST

MOTORING SAFETY
IAN BREACH

HUMANIST HUMANIST

THE DECLINE OF MAGIC
PETER CADOGAN

HUMANIST HUMANIST

INTELLECTUALS & TREASON
ROGER MANVELL

HUMANIST HUMANIST

NEW CAMPAIGN AGAINST PORN
BENN LEVY

HUMANIST HUMANIST

POEMS

PHILIP LARKIN & DONALD DAVIE

HUMANIST HUMANIST

Research: QUASI-RATIONAL QUARRELLIN

HUMANIS

Links and Full-text

1. Full-text [articles](#) from the Rationalist Annual by the prolific biologist, popular science writer and communist, [J. B. S. Haldane](#). Some titles:

- Why I am a Materialist
- The Laws of Nature [probabilistic theories in physics and genetics, Laplace's Demon]
- The Limitations of Rationalism [cognitive limitations and biases -- see below]
- A Rationalist Approach to the Problem of Sexual Relations
- On Being Finite [reflections on death and meaning in life -- see below]

2. Full-text [article](#) by Ernest Gellner "On Being Wrong" [epistemic corrigibility]

3. Wikipedia on the (British) [Rationalist Association](#), the [Indian Rationalist Association](#), and the [Rationalist International](#).

4. [Archive](#) of the Rationalist Annual and New Humanist (pay-walled).

Some Random Quotes

Yet it seemed to me that we have greater control over the material than the immaterial; it being easier, for instance, to regulate diet than to command success [see [Scott Alexander](#)]; and that, this being so, we must learn to call nothing common nor unclean, but consider a careful investigation into the influences of food, medicine, and climate as beneath the notice neither of philosopher nor moralist

...
'Of all the great branches of human knowledge, medicine is that in which the accomplished results are most obviously imperfect and provisional. The medicine of inhalation is still in its infancy; and yet it is by inhalation that nature produces most of her diseases, and effects most of her cures'

From "Mind as controlled by matter" by Constance Plumptre (1896)

More serious is my own and other people's limited reasoning power. I simply cannot grasp a sufficiently complicated argument. The limitations of different intellects vary. Every educated person can factorize 35 in his or her head. I can factorize such numbers as 11,009; a few people can factorize six-figure numbers, however large their factors, Nobody can factorize twenty-figure numbers with large factors. This kind of thing can be done with a specially designed machine, and machines can already vastly extend the scope of our reasoning power in the field of mathematics. I believe that we are only at the very beginning of the use of mechanical aids to reasoning, and that they will be as important as mechanical supplements to our senses, such as microscopes. But they will only take us a certain distance.

From "The Limitations of Rationalism" by J. B. S. Haldane (1947 -- [link](#))

I have tried morphine, heroin, and *bhang* and *ganja* (hemp prepared for eating and smoking). The alterations of my consciousness due to these drugs were trivial compared with those produced in the course of my work [as an evolutionary biologist]. I once dreamed that I was reading a life of Christ written and illustrated by Edward Lear. But I can only remember Pontius Pilate's moustache. If you want a dream as original as that, don't take opium, but eat sixty grams of hexahydrated strontium chloride [not medical advice -- OE]. I have had some of the standard adventurous experiences such as being pulled out of a crevasse in a glacier, and more which are unusual. For example, I was one of the first two people to pass forty-eight hours in a miniature submarine, and one of the

first few to get out of one under water. I doubt whether, given my psychological make-up, I should have found many greater thrills in a hundred lives. So when the angel with the darker drink at last shall find me by the river's brink, and offering his cup, invite my soul forth to my lips to quaff, I shall not shrink.

From “On Being Finite” by J. B. S. Haldane (1965 -- [link](#))

Bonus

Past presidents of the Rationalist Association include Bertrand Russell, the [founder](#) of the Bletchley Park estate, the [co-founder](#) of the London School of Economics, and my dad's [ex-father-in-law](#).

Here's an effusive advert for a book on “The Science of Life” by H. G. Wells, Julian Huxley and G. P. Wells:

The Literary Guide

AND RATIONALIST REVIEW

[ESTABLISHED 1885]

No. 438 (NEW SERIES)

DECEMBER, 1932

MONTHLY : 3d.

FOR WRITERS, CONVERSATIONALISTS, SPEAKERS, & THINKERS

STRAINGER than any flights of fiction fancy are the revelations in THE SCIENCE OF LIFE. Here at last is a complete survey of the whole of this vast subject, not in the dry-as-dust style of the Specialist in Science, but with all the combined Wells and Huxley powers of clear and cogent presentation.

You Are Living Your Life in the midst of a mighty Sea of Science Secrets. Learn about them. The knowledge you can gain from THE SCIENCE OF LIFE will give you a new and profound understanding

—a new and infinitely greater and more interesting outlook on life.

Your Conversation and Writing will be enriched from a vast fund of Science. Your Knowledge will extend to new horizons. Your Mind will be stored with an infinitely varied wealth of fact-material and a host of thought-challenging and inspiring ideas.

Post The Coupon and receive gratis and post-free a large Illustrated Album-Portfolio more fully describing THE SCIENCE OF LIFE.



H. G. WELLS

One of the world's supreme living examples of Scientific Knowledge and Literary expression. This great Scientific Survey is a Masterpiece of Knowledge.

SOME OF THE CONTENTS OF FREE PORTFOLIO

- The Greatest of All Mysteries—LIFE
- Its Origins, Its Evolutions
- Life's Myriad Forms—Visible and Invisible
- A Telescope and Microscope for Your Mind
- The Glands of Life (Illustrated)
- Your Wonderful Nerves (Illustrated)
- Human Dwarfs and Giants (Illustrated)
- Nature's Extraordinary Births (Illustrated)
- Secrets of Sex and Reproduction (Illustrated)
- The Nourishment of the Body
- Problems of Human Health and Disease
- Is Man at the Threshold of a New Spiritual Era? (Amazing Spirit Photographs)
- Many Unique X-Ray, Microphoto, and other Illustrations Never Before Published

Amazingly Informing!

The Greatest Work of Knowledge of To-day

THE SCIENCE OF LIFE

By H. G. WELLS, JULIAN HUXLEY, and G. P. WELLS

In 3 Large Library Volumes

Bound in Art Canvas with Gilt Lettering and Burnished Top Edges.
31 Plates in Full Colour, 272 Drawings, and 712 Photographic
and other Revealing Illustrations.

IMMEDIATE

26

For only

DELIVERY

1st Subscription

The prices quoted herein do not apply to South African or Australian readers, who should apply for full particulars to the Standard Agency, 10, Gloucester Place, London, W.C.1. Readers in India, Burma, and Ceylon should write to The Standard Literature Co., Ltd., 181 Old Court House St., Calcutta. Other purchasers abroad should write to London for particulars.

To every THINKING person this Master Work dealing with the Secrets of Life will prove most astonishingly challenging, informing, and inspiring. This great and up-to-date Master-piece of Science is already in the Libraries of thousands of Authors, Scientists, Lecturers, Teachers, and Thinkers.

Simply write your NAME and ADDRESS on the Coupon and POST IT NOW in 4d. stamped UNSEALED Envelope.

POST THIS COUPON TO-DAY



"LITERARY GUIDE" COUPON

FOR "SCIENCE OF LIFE" PORTFOLIO

FREE

To THE WAVERLEY BOOK CO. LTD.
96-97 Farringdon Street, London, E.C.4

SIRS.—Please post me a FREE COPY of the PORTFOLIO (Illustrated) describing THE SCIENCE OF LIFE, and its over 1,000 Illustrations in Colour and Monochrome, 3 well-bound Quarto Library Volumes. It is understood that I may have Immediate Delivery of the complete work (carriage paid) on a first subscription of 26/- only.

NAME.....

ADDRESS.....

L.G.L.1932

Preprint is out! 100,000 lumens to treat seasonal affective disorder

Let's give people with winter depression (seasonal affective disorder, SAD) LOTS OF LIGHT and see what happens!

Our preprint is out now for our paper "100,000 lumens to treat seasonal affective disorder: A proof of concept RCT of bright, whole-room, all-day (BROAD) light therapy"! We have sent it to a number of professors working in that area and have received very encouraging and helpful feedback, which has made me even more excited about continuing this research.

→ Paper: <https://medrxiv.org/cgi/content/short/2021.10.29.21265530v1>, short summary Twitter thread:
<https://twitter.com/FabienneSand/status/1457745472773296128>

Jan Brauner and I are very thankful to the LessWrong/EA communities, which have inspired this first study (there will be more) and through which we have found funding. In particular, thank you Eliezer Yudkowsky for helping us find funding and for inspiring the study with Inadequate Equilibria, David Chapman for inspiring us with [these two](#) posts in the Meaningness blog, Raemon for inspiring us with [this](#) LessWrong post and everyone who discussed with us setups they have tried. <3

Split and Commit

This is an essay describing a basic sanity-inducing mental movement that I use approximately ten times per week, and suspect other people would benefit from adopting (and regularly reminding each other to do). I've talked about it [elsewhere](#), but until now it didn't have its own linkable reference post.

P1: Things are often not what they seem.

This can be because the seeming itself is broken, e.g. the progression that goes:

"Well, here I am in 1700's America and it sure seems to me that these black folk are fundamentally intellectually and morally and spiritually inferior to white folk" or "Well, here I am in 1800's America and it sure seems to me that these women are constitutionally incapable of holding political office" or "Well, here I am in 1900's America and it sure seems to me that these gay men are an active threat to everyone's safety, including our women (somehow) and children" or "Well, here I am in 2000's America and it sure seems to me that [REDACTED]"

[REDACTED]."

(Or "Well, here I am inventing physics and it sure seems to me that force consistently and exactly equals mass multiplied by acceleration.")

It can also be because the seeming is generally correct, but there exists variance. If 95% of the marbles in a bag are red, and 5% of the marbles in the bag are green, it's correct to guess that the next *randomly selected marble* will be red, because betting on red makes you the least likely to be wrong. But that doesn't mean that it *can't* be green, or that you should be *shocked* if it's green. If you bet on red the whole way through the bag, you *know* you'll be wrong five percent of the time.

P2: Every observation is evidence in favor of more than one hypothesis.

If my friend tells me my hair looks good, this is an update in favor of my hair looking good *and* an update in favor of my hair not looking good but my friend lying to me for various reasons (kindness, prank). In order to know which update is *larger*, you need some other information about what kind of person my friend is, and what our relationship is like.

If I see some breaking news about some person having maybe done something horrible, this is an update in favor of that person having done something horrible, *and* an update in favor of that person being the target of a smear campaign/conspiracy, *and* an update in favor of the presented claim being technically correct but there being a ton of other relevant context and facts that will substantially change the gestalt of the story. In order to know which update is *larger*, you need some other information, etc.

(Many of my readers will be thinking "this is just Bayes," and yep, it's just Bayes.)

The upshot of these two premises is that it's wise to have *at least* two active hypotheses running at all times, for any question of import.

By default, it seems to me that most humans have only one, i.e. not a hypothesis at all but a *singular belief*.

This is true (in my experience) even of the type of person who is acutely aware of their own fallibility, and who acknowledges that they are occasionally (or often) mistaken. Even people who are savvy enough to try to put a number (like "85% confident") on it, such that they can track their calibration over time, tend to leave the remaining swath of possibility un- or under-specified.

They will admit that they *might be wrong* in some vague fashion about [free speech, COVID policy, the president, existential risk from artificial intelligence, Kyle Rittenhouse, universal basic income, Leverage Research, the importance of religion, what the person who wrote that comment was obviously *really* saying, never mind their actual words], etc., but that doesn't stop them from *having only a single real guess* in many situations where that seems (to me) to be wildly premature, and an open invitation to all sorts of known and problematic bias.

There's a huge difference between [having an answer which you are virtuously prepared to abandon, if forced] and [keeping two or more distinct possibilities firmly in focus, even as you track that one of them is substantially more likely than the others].

There's a difference in how that feels, and there's a difference in how it influences one's reactions to new and relevant information. Choosing a single possible world-state and then looking to see if it's compatible with the available evidence is *very different* from looking at multiple possible world-states and then asking yourself what sorts of evidence would rule each one out.

To be clear: it's true that one (often, regrettably) has no choice but to do the equivalent of [placing an unambiguous bet] when it comes to *taking actions*.

Sometimes, you simply have to behave as though the most likely outcome *is* what's going to happen—to choose [actions that will pay out if the marble is red, and cost you if the marble is green] over [actions that will pay out if the marble is green, and cost you if the marble is red]. There are many situations where we cannot afford to sit back and do nothing while we wait for more information to come in, and in many of those situations we have to pick a single exclusive strategy and run with it. You can't always hedge.

But the fact that one must (often, unfortunately) take singular *action* doesn't mean that one can't hold nuanced *beliefs*. You can put your money on red without losing track of the true fact that the next marble out of the bag could easily be green.

Split and Commit

Here, then, is the recommendation:

When you encounter [evidence] that *sure looks to you* like it implies [X], then rather than simply switching into "evaluate [X]" mode, you *split and commit*.

By "split," I mean that you explicitly ask yourself *both*:

"What kind of world contains both [evidence] and [X]?"

and also:

"What kind of world contains both [evidence] and [not-X]?"

Don't just focus on the world where things are what they seem to be, at first blush.

Feel free to notice that one possibility seems pretty darn likely, but hold yourself to the standard of seriously and concretely considering *at least one other possibility*.

"If it turns out that this *isn't* what it looks like, what's the next most likely story that's still consistent with [evidence]?"

And by "commit," I mean that you choose a preliminary reasonable-to-you response in *each* of those possible worlds.

"If this is indeed what it looks like, then I should probably do something like [A]. If it's not, though, then [A] would be bad/counterproductive, and I should instead respond with something like [B]."

This commitment doesn't have to be public. It doesn't have to be set in stone. It can be conditional, depending on the various possible states of your next piece of evidence.

The key thing is simply that it *exist at all*. That you set aside the additional thirty seconds it takes to specifically and concretely dignify the possibility that things are not what they currently appear to be, and make a rough draft of what right action looks like, in that world. That you do this habitually, so that in those times when they *aren't* what they seemed at first glance, you're primed to notice, and ready to respond.

Or, if I might embed my tongue firmly in my cheek: if running this algorithm seems to you like a dumb or not-worth-it idea, then fair enough, but...what ought you do in the world where it just *looks* dumb, and actually isn't?

Appendix

The following are some responses from people who encountered the split-and-commit tool in earlier essays and had concrete things to say about it.

Rob Bensinger:

A benefit of split-and-commit I'm surprised wasn't high on your list: people often want to hedge their bets and pick stances and policies that internally feel justified/OK regardless of what the outcome is—they like being able to strategically switch between 'that looks bad' and 'that *is* bad'. Split-and-commit makes it easier to catch yourself doing this and discourage anyone from doing it.

Irena Kotíková:

One of my favourite concepts of all time. Especially because I often notice a ton of resistance to keeping the commitment once the split happens.

Marcello Herreshoff:

So the way I see it, it feels more like split and commit has a minimum of *three* plans. You need the third plan to tell you what experiment to do to figure out which of the two worlds under consideration you live in. Otherwise tomorrow could easily look like "yep; it still seems like things are as they seem, time to execute plan one!"

Logan Strohl:

Immediately after I finished reading this, I practiced the very first step in a training progression for "split and commit" on a walk from my house to the library.

Here's the exercise I did:

While walking, my attention will happen to land on things. When it does, I'll run through the following structure: It looks like x. It might instead be that y. If it's what it looks like, I will p. If y, I will q.

Some examples I happen to remember:

- It looks like that's a building. It might instead be an alligator. If it's what it looks like, I'll just walk by it. If it's an alligator, I'll go get my neighbor to test whether there is in fact a fucking alligator.
- Seeing that guy from the back, it looks like he's smoking a cigarette. It might be that he's not smoking a cigarette. If he is, I'll walk by him and hold my breath. If he's not, I'll walk by him and breathe normally.
- That looks like a magnolia tree. It might be a tree I'm unfamiliar with. If it's a magnolia tree, I'll expect to keep seeing things I associate with magnolia trees if I keep inspecting it. If it's a tree I'm unfamiliar with, I'll expect to encounter things I don't associate with magnolia trees if I keep inspecting it.

Notes on my experience of the walk:

- Often when I identify an alternative interpretation of some observation, the course of action I commit to is the same in either case, but I come away with a feeling of having learned something anyway.
- It looks to me like most reasonable responses are expectations of future experiences, rather than physical interventions or policy changes. It may be that this changes when the motion "split and commit" is taken only when I encounter an appropriate trigger, rather than arbitrarily. (I notice myself automatically deciding on courses of action given either state of affairs; yes good.)
- The one-second version of split and commit involves going back-and-forth two times. The first time you feel the thing you're seeing from your default perspective, and then flip over to a perspective where you feel its meaning from a different perspective. The second time, you occupy the first perspective while imagining a course of action to take in response, then flip over to the other perspective and imagine an appropriate course of action from there. I started doing the one second version after about seven

minutes of practice, which was about three minutes after it became available. I stuck to the slow version for the extra three minutes to make sure I knew what I was doing.

- There are a lot of alternative interpretations of the same observations, although there may only be a small number that explain the observation about equally well. There is often a moment where I must choose whether to keep generating interpretations. It seems to me from my experience practicing this so far that most of the benefit most of the time comes from identifying a single alternative interpretation, followed by plans of action for each interpretation (although in fact bothering to identify an alternative interpretation is all by itself most of the thing). I have a feeling that there are some kinds of situations where it is wiser to generate several interpretations before committing to courses of action; I will file this under “followup study”, and continue to focus on “split and commit”.

As usual, humans are difficult and complicated and I’m glad I began to practice mostly in their absence, even though this was presented [in its original context] as primarily a social skill. I think focusing on humans should probably be part three of the training progression.

(Part two of the training progression should be to identify and train the triggers for “split and commit”; obviously splitting and committing for absolutely anything my attention happens upon is only useful if I want a rapid-fire training session for the motion itself. The motion is best taken in response to certain kinds of experiences, and the next thing I need to know are *which* experiences indicate that it’s an especially good time to split and commit.)

Comments on Carlsmith's "Is power-seeking AI an existential risk?"

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

The following are some comments I gave on Open Philanthropy Senior Research Analyst Joe Carlsmith's Apr. 2021 "[Is power-seeking AI an existential risk?](#)", published with permission and lightly edited. Joe replied; his comments are included inline. I gave a few quick replies in response, that I didn't want to worry about cleaning up; Rob Bensinger has summarized a few of them and those have also been added inline.

I think Joe Carlsmith's report is clear, extensive, and well-reasoned. I also agree with his conclusion, that there's at least a 5% chance of catastrophic risk from AI by 2070. In fact, I think that number is much too low. I'll now attempt to pinpoint areas of disagreement I have with Joe, and put forth some counterarguments to Joe's position.

Warning: this is going to be a bit quick-and-dirty, and written in a colloquial tongue.

I'll start by addressing the object-level disagreements, and then I'll give a few critiques of the argument style.

On the object level, let's look at Joe's "shorter negative" breakdown of his argument in the appendix:

Shorter negative:

By 2070:

1. It will become possible and financially feasible to build APS AI systems.

65%

2. It will much more difficult to build APS AI systems that would be practically PS-aligned if deployed than to build APS systems that would be practically PS-misaligned if deployed, but which are at least superficially attractive to deploy anyway | 1.

35%

3. Deployed, practically PS-misaligned systems will disempower humans at a scale that constitutes existential catastrophe | 1-2.

20%

Implied probability of existential catastrophe from scenarios where all three premises are true: ~5%

My odds, for contrast, are around 85%, 95%, and 95%, for an implied 77% chance of catastrophe from these three premises, with most of our survival probability coming from "we have more time than I expect". These numbers in fact seem a bit too low to me, likely because in giving these very quick-and-dirty estimates I failed to account properly for the multi-stage fallacy (more on that later), and because I have some additional probability on catastrophe from scenarios that don't quite satisfy all three of these conjuncts. But the difference between 5% and 77% is stark enough to imply significant object-level disagreement, and so let's focus on that first, without worrying too much about the degree.

"we have more time than I expect"

Joe Carlsmith: I'd be curious how much your numbers would change if we conditioned on AGI, but after 2070.

[*Partial summary of Nate's reply:* Nate would give us much better odds if AGI came after 2070.]

I have some additional probability on catastrophe from scenarios that don't quite satisfy all three of these conjuncts

Joe Carlsmith: Would be curious to hear more about these scenarios. The main ones salient to me are "we might see unintentional deployment of practically PS-misaligned APS systems even if they aren't superficially attractive to deploy" and "practically PS-misaligned APS systems might be developed and deployed even absent strong incentives to develop them (for example, simply for the sake of scientific curiosity)".

Maybe also cases where alignment is easy but we mess up anyway.

[*Partial summary of Nate's reply:* Mostly "we might see unintentional deployment of practically PS-misaligned APS systems even if they aren't superficially attractive to deploy", plus the general category of weird and surprising violations of some clause in Joe's conditions.]

Background

Before I dive into specific disagreements, a bit of background on my model of the world. Note that I'm not trying to make a large conjunctive argument here, these are just a bunch of background things that seem to be roughly true-ish to me and that inform where I'm coming from in my following criticisms. Note also that Joe mentioned / acknowledged many of these points (or related points).

1. Much of my strategic thinking about advanced AI systems revolves around the notion of "decisive strategic advantage" -- it's all well and good if your AI can make a bunch of money on the stock market, but one capable of wiping out the entirety of humanity in an afternoon is a different story altogether.

- For Team Earth, what matters is the ability to carry out a much more limited "pivotal action", that ends the acute risk period, eg by deploying some AI system in a way that decisively prevents any other AI system from wiping out the planet, while passing control to some process that can make sure that the future is rad and that the cosmic endowment is well-spent.
- By contrast, an AI system with no love (nor hatred!) for humanity has a much wider range of less delicate and more destructive actions available, such as manufacturing nanomachines that silently proliferate and disperse until they have entered the bloodstream of literally every human, at which point they kill all humans simultaneously.

2. The bottleneck on decisive strategic advantages is very likely cognition (of a deep and high-quality variety).

- The challenge of building the aforementioned nanomachines is very likely bottlenecked on cognition alone. (Ribosomes exist, and look sufficiently general to open the whole domain to any mind with sufficient mastery of protein folding, and are abundant.)
- In the modern world, significant amounts of infrastructure can be deployed with just an internet connection -- currency can be attained anonymously, humans can be hired to carry out various physical tasks (such as RNA synthesis) without needing to meet in person first, etc.
- The laws of physics have shown themselves to be "full of exploitable hacks" (such as the harnessing of electricity to power lights in every home at night, or nuclear fission to release large amounts of energy from matter, or great feats of molecular-precision engineering for which trees and viruses provide a lower-bound).

3. The abilities of a cognitive system likely scale non-continuously with the depth and quality of the cognitions.

- For instance, if you can understand protein folding well enough to get 90% through the reasoning of how your nanomachines will operate in the real world, that doesn't let you build nanomachines that have 90% of the impact of ones that are successfully built to carry out a particular purpose.
- I expect I could do a lot with 100,000 trained-software-engineer-hours, that I cannot do with 1,000,000 six-year-old hours.

Joe Carlsmith: Does the notion of "recursive self-improvement," or of broader feedback loops where AI capabilities (and their economic outputs) are reinvested in improving AI hardware/software, play an important role in your backdrop picture? Open I has been doing some work to understand various models of this sort of dynamic, and it's been pretty central to traditional stories about dramatic take-offs/DSAs, but you don't mention it here, and below you say that you don't put a lot of weight on "also things might speed up a bunch when we get close & can build tools that help us speed up."

[*Partial summary of Nate's reply:* Once dangerous AGI systems exist, if you lose control of one, it may well get scarier than you expect surprisingly quickly because of things like recursive self-improvement.]

In worlds where AGI is developed in an ML paradigm like the current one, Nate's models do not strongly rely on (nor strongly rule out) the possibility that AGI development is sped up by the (pre-AGI) fruits of AI progress.]

if you can understand protein folding well enough to get 90% through the reasoning of how your nanomachines will operate in the real world, that doesn't let you build nanomachines that have 90% of the impact of ones that are successfully built to carry out a particular purpose.

Joe Carlsmith: Not sure about this -- feels plausible to me that being able to "almost" build nano-tech (at some specified level of quality) allows you do some pretty cool stuff. Do you think there's some important difference here between e.g. nano-tech and e.g. microsoft word? It's true that there's a big difference between a 90% functional MS-word (depending on what 10% we remove), and a 100% functional one -- but you still see versions of MS-word steadily increasing in quality over time, rather than a discontinuous jump. You could make similar arguments about e.g. planes. In some sense there's a discontinuous difference between "can fly" and "can't fly," but you still start with pretty crappy planes (from AI impacts, <https://aiimpacts.org/discontinuous-progress-in-history-an-update/>: "Powered heavier-than-air flight got started in 1903, but at first planes only traveled hundreds of feet, and it took time to expand that to the 1600 or so miles needed to cross the Atlantic in one hop"), which then get better over time.

One can claim that specific technologies will especially amenable to discontinuities; but it felt like, here, you wanted to make a broader point about cognition per se.

[*Partial summary of Nate's reply:* Getting 90% of the way through the process of figuring out a design for a protein factory, that you were hoping to use to build a nanofactory that you were hoping to use to build nanomachines that could execute a pivotal act, would let you do a lot of cool things. But it wouldn't let you 90%-save the world.]

I expect I could do a lot with 100,000 trained-software-engineer-hours, that I cannot do with 1,000,000 six-year-old hours.

Joe Carlsmith: I find this sort of example more convincing. In particular, it seems plausible to me that as a matter of empirical fact, there's a big gap in the quality of cognition in e.g. a $1e15$ parameter model and a $1e16$ parameter model, such that we move sufficiently fast across the gap that we see dramatic increases in capability across a given unit of scaling.

Joe Carlsmith: I'm most used to thinking about claims like 1-3 in the context of a picture where, shortly after we start developing better-than-human systems (or perhaps, shortly after we reach some-harder-to-pinpoint capability threshold?), there will be a single AI system that can kill all humans and take over the world fairly easily, even despite the level of resistance/competition coming from other parts of the world also scaling up their own AI capabilities, trying to defend themselves, trying to stabilize the situation, etc. Does that sound like a reasonable description of your mainline scenario?

If so, I'd be curious to break down your levels of optimism/pessimism in worlds where we vary some of these assumptions -- e.g., we make take-off slower, we make DSAs relative to the rest of the world harder to get, we make the gap between "AGI/TAI" and "can build killer nano-tech using mostly its mind and readily available tools" bigger, etc. I expect that stuff in this vicinity is an important source of disagreement.

[*Partial summary of Nate's reply:* The case for DSAs being available and cognition-bound seems pretty open-and-shut to Nate. He considers it plausible that there's a crux (which Joe might be gesturing at) around "there's a difference between the narrow stuff and the general stuff, and the general stuff is going to hit the Earth like a hammer".]

Ok, great. Next, a caveat:

Joe focuses on systems with the following three properties:

- *Advanced capability:* they outperform the best humans on some set of tasks which when performed at advanced levels grant significant power in today's world (tasks like scientific research, business/military/political strategy, engineering, and persuasion/manipulation).
- *Agentic planning:* they make and execute plans, in pursuit of objectives, on the basis of models of the world.
- *Strategic awareness:* the models they use in making plans represent with reasonable accuracy the causal upshot of gaining and maintaining power over humans and the real-world environment.

I have my own concept of "AI systems worth worrying about", that does not quite fit cleanly into this definition. I am basically just going to run with my own concept, instead of trying to contort my thinking around Joe's. I expect this to mostly go fine, as my sense is that Joe was, with his definitions, mostly trying to wave in the direction of the cluster of things that my concept also covers. That said, some of our apparent disagreement might come from the fact that I'm using my concept, and he's using his narrower concept.

My broader concept, for the record, is something more like "AI systems that can carry out long and deep cognitive chains, sufficient for accessing the Scary Stuff", where the "AI systems" part connotes that it's more like an alien mind than like an extension of a human's will (which rules out, eg, whole brain emulation or genetically engineered super-smart human beings).

Joe Carlsmith: I'm assuming the Scary Stuff is centrally stuff like nano-tech? I'd be interested in hearing a few more examples -- when I've tried to think through DSA-ish scenarios, stuff like nano-tech and bio-tech often seems uniquely scary, especially when accompanied by ~unlimited hacking ability, but that if we cut that stuff out, the stories get harder to tell (though obviously, "things we can't think of right now" is a big bucket of scary as well -- and plausibly the biggest by far). For example, think stories of DSA via persuasion, general economic productivity, seizing control of conventional weapons/infrastructure, etc. get harder to tell, especially if you don't assume unlimited hacking ability.

[*Partial summary of Nate's reply:* On Nate's view, unlimited hacking ability, nanotech, biotech, directly hacking human minds, and physics exploits not known to us are easily sufficient to carry the argument.]

Joe Carlsmith: For my own purposes, I wouldn't want to build "not an extension of a human's will" into the "AI systems worth worrying about" concept, because I want it to be an open question whether, once we can build such systems, they'll be aligned. And fwiw, i think you could have relevantly misaligned WBEs, genetically engineered super-humans, etc (and more generally, seems possible to have e.g. perfect butlers that are still "alien minds" in a recognizable sense, as opposed to something like "tools"). Not sure that much rides on this though.

Some places where I suspect our concepts differ, in ways that are perhaps relevant:

1. I don't think strategic awareness matters all that much, from the perspective of Team Earth. Or in other words: the AIs that are likely to kill us, are likely strategically aware, but if we successfully deploy AIs in a way that ends the acute risk period, I think it's about even-odds as to whether they're strategically aware, and the question doesn't factor much into my capability estimations.

Joe Carlsmith: I'm not sure there's actually any disagreement here, except insofar as you see strategic awareness as genuinely optional in the misaligned systems that end up killing us, in which case my response is: it seems like strategic awareness is pretty crucial for the instrumental convergence argument to work (indeed, the concept is crafted to play the necessary role in the instrumental convergence arg). I agree that we could maybe use non-strategically aware systems to end the acute risk period -- but "AI systems worth worrying about" is different from "AI systems that would be sufficient for ending the acute risk period."

2. I suspect I think "agentic planning" is a much weaker condition than Joe does.

- For instance: I say things like "There are hugely many ways to reduce your predictive error when thinking about protein folding, only a few of which matter for the ability of the resulting nanosystem to carry out some desired behavior in the real world; any AI system that is capable of designing a system that carries out a specific behavior in the real world, is *somewhat* focusing its cognition on the questions of consequential behavior, in an action I'd call 'planning' 'in pursuit of an objective'."
- For instance: I say things like "I care little about whether you think of the system as having unfettered access to a motor system, or as only ever producing schematics for nanomachines that you synthesize yourself; anything that gives the AI significant causal influence over the world looks to me like it is allowed to 'act' in ways that 'execute plans on the world'."

There are hugely many ways to reduce your predictive error when thinking about protein folding, only a few of which matter for the ability of the resulting nanosystem to carry out some desired behavior in the real world

Joe Carlsmith: Not sure I've understood this point. What's the difference here between the methods of reducing your predictive error that "matter" to the nanosystem's functioning, and the ones that don't? Maybe an example would help.

[*Partial summary of Nate's reply:* Suppose that you're trying to design a protein factory so that you can build a nanofactory. This is a tricky engineering task. As a subproblem, you want to develop the ability to predict protein interactions, with the intent of being able to use your models when designing open-air nanotech.

You train a large model on protein interactions, using the mean-square distance of each modeled-atom-position from its corresponding real-world position. Most atoms, however, are moving rapidly in the middle of some enormous enzyme, while a couple are on the binding site.

Among the ones on the binding site, all that matters is figuring out whether the enzyme reacts with the catalyst you are considering using to carry out some specific function in your protein factory. Your model, however, doesn't care about that; it is minimizing the mean-square distance of all atoms, and therefore spends all day learning about the behaviors of atoms in the middle of the enzyme.

The world is extraordinarily detailed. On Nate's view, if we want a plan that works in real life, we have to concentrate cognition on the relevant questions. As Nate sees it, asking for a "non-consequentialist but working plan" is a bit like asking for a "way of drawing an accurate picture of a maze without observing it".]

any AI system that is capable of designing a system that carries out a specific behavior in the real world, is *somewhat* focusing its cognition on the questions of consequential behavior, in an action I'd call 'planning' 'in pursuit of an objective'."

Joe Carlsmith: I'm pretty sympathetic to this, provided that the "planning" in question allows for the type of predictions about the system's behavior required for the instrumental convergence argument to go through (the instrumental convergence arg is ultimately why I care about agentic planning and strategic awareness).

That is, I think the question is whether the type of planning strictly required to e.g. build a functioning nano-system supports the instrumental convergence arg (for example, can you build nano-tech in a way that is intuitively 'sphexish'? If not, is this centrally about the complexity of the task? Evolution can build squirrels, for example, that hide their nuts in trees for the winter, but that don't take other actions that might help them survive for the winter (or whatever; I don't actually know the squirrel science); we can train AlphaStar to execute cool plans in a StarCraft environment, but in a way that wouldn't generalize to strategic action in other contexts; and so on.)

[*Partial summary of Nate's reply:* Nate agrees that if there's a sphexish way to build world-saving nanosystems, then this should immediately be the top priority, and would be the best way to save the world (that's currently known to us). Nate doesn't predict that this is feasible, but it is on his list of the least-unlikely ways things could turn out well, out of the paths Nate can currently name in advance. (Most of Nate's hope for the future comes from some other surprise occurring that he hasn't already thought of.)]

unfettered access to a motor system

Joe Carlsmith: Fwiw, this isn't necessary for agentic planning in my sense: see e.g. the bullet points in 2.1.2: <https://docs.google.com/document/d/1smal1lagHHcrhoi6ohdq3TYIZv0eNWWZMPEy8C8byYg/edit#heading=h.70ean6ha5tu6>.

producing schematics for nanomachines that you synthesize yourself

Joe Carlsmith: I do think there's an important question, though, about how exactly those schematics were optimized. That is, you can imagine an AI (system A) with the myopic goal of producing, on this single pass through the network, a plan for a nano-system that would receive high ratings according to some other system's (system B) predictions about whether that plan would result in a nanosystem that works. System A has an opportunity to exert causal influence on the world, yes, via these planned nano-systems getting built; but the nano-systems themselves aren't optimized directly for implementing some world-take-over scheme. Rather, they are optimized such that the plans they get built from receive high ratings from system B.

[*Partial summary of Nate's reply:* Nate's model says that if system A can produce plans that would work, it's at least scary-

[adjacent. He also believes that in assuming that we successfully got a scary system to have a single myopic goal, we are assuming that we have already solved the hard part of the alignment problem.]

3. I suspect Joe has some concept of advanced AI systems that are not "agentic planners", and that have a different catastrophe-profile.

- The most obvious example people in our neck of the woods often discuss is systems that are "deferential to humans" in some deep and persistent way.
- My basic take on this question is "that's doubtful (that humanity will be able to pull off such a thing in the relevant timeframes)". It seems to me that making a system "deferential all the way down" would require a huge feat of mastery of AI internals that we're nowhere close to.
- Paul Christiano occasionally floats proposals of (what looks to me like) deferential cognitive systems that are too incapable to be scary, being composed into highly capable cognitive systems that inherit a deference property from their parts. (Paul might not endorse this gloss.) I basically expect the cognition to not compose to something capable, and insofar as it does I basically expect it not to inherit the deference property, and so I have little optimism for such approaches. But it's possible that Joe does, and that as such, the second bullet point above is doing a bunch of work for him that it's not doing for me.

The most obvious example people in our neck of the woods often discuss is systems that are "deferential to humans" in some deep and persistent way.

Joe Carlsmith: I wouldn't be thinking of "deferential systems" as non-APS. A perfect butler, for example, makes plans in pursuit of objectives, with strategic awareness. Same with a perfect AI personal assistant.

To me, the most salient "no APS systems" scenarios involve intuitively "sphexish" systems whose 'planning' is sufficiently brittle/model-free/domain-limited that it doesn't generalize in response to power-seeking type incentives. I think of Drexler's "CAIS" vision as embodying something like this vibe. When I imagine this, I imagine the way I feel about image classifiers (e.g. "in a fairly structural way, this isn't the type of system we need to be worrying about" -- though this need not hold if the system is scaled up to arbitrary degrees) turning out to apply to many more systems than I'm inclined to expect by default.

I also have some probability on "this whole vaguely-rational-agent-EU-maximizer way of thinking about AI systems is misguided somehow." The report isn't explicitly framed in terms of maximizing utility functions, but I expect that similar background concepts/assumptions are sneaking in at various points, and I think this may well just result in confusion/mis-prediction at levels that can be difficult to anticipate/conceptualize.

it's possible that Joe does, and that as such, the second bullet point above is doing a bunch of work for him that it's not doing for me.

Joe Carlsmith: Optimism/pessimism about Paul's approach isn't playing a role in the probability I assign to "timelines" or "incentives to build APS systems". I think of Paul's work as centrally about building aligned APS systems, rather than as trying to avoid APS-ness. Related to the "perfect butlers are still agentic planners" point above.

Timelines

Now, on to critiquing Joe's probabilities:

Re (1), on timelines, observe that my timelines are significantly more aggressive than Joe's. I'm not sure where the disagreement lies, but I can list a handful of things that drive my probability of AGI-in-the-next-half-century-or-so above 80%:

1. 50 years ago was 1970. The gap between AI systems then and AI systems now seems pretty plausibly greater than the remaining gap, even before accounting the recent dramatic increase in the rate of progress, and potential future increases in rate-of-progress as it starts to feel within-grasp.

2. I observe that, 15 years ago, everyone was saying AGI is far off because of what it couldn't do -- basic image recognition, go, starcraft, winograd schemas, simple programming tasks. But basically all that has fallen. The gap between us and AGI is made mostly of intangibles. (Computer programming that is Actually Good? Theorem proving? Sure, but on my model, "good" versions of those are a hair's breadth away from full AGI already. And the fact that I need to clarify that "bad" versions don't count, speaks to my point that the only barriers people can name right now are intangibles.) That's a very uncomfortable place to be!

3. When I look at the history of invention, and the various anecdotes about the Wright brothers and Enrico Fermi, I get an impression that, when a technology is pretty close, the world looks a lot like how our world looks.

- Of course, the trick is that when a technology is a little far, the world might also look pretty similar!
- Though when a technology is very far, the world *does* look different -- it looks like experts pointing to specific technical hurdles. We exited that regime a few years ago.

4. Summarizing the above two points, I suspect that I'm in more-or-less the "penultimate epistemic state" on AGI timelines: I don't know of a project that seems like they're right on the brink; that would put me in the "final epistemic state" of thinking AGI is imminent. But I'm in the second-to-last epistemic state, where I wouldn't feel all that shocked to learn that some group has reached the brink. Maybe I won't get that call for 10 years! Or 20! But it could also be 2, and I wouldn't get to be indignant with

reality. I wouldn't get to say "but all the following things should have happened first, before I made that observation!". Those things have happened. I have made those observations.

5. It seems to me that the Cotra-style compute-based model provides pretty conservative estimates. For one thing, I don't expect to need human-level compute to get human-level intelligence, and for another I think there's a decent chance that insight and innovation have a big role to play, especially on 50 year timescales.

6. There has been a lot of AI progress recently. When I tried to adjust my beliefs so that I was *positively* surprised by AI progress just about as often as I was *negatively* surprised by AI progress, I ended up expecting a bunch of progress.

There are other arguments against short timelines that I don't find very compelling (eg "society might collapse"), there are other arguments in favor that I don't find super compelling (eg "also things might speed up a bunch when we get close & can build tools that help us speed up"). \shrug.

My overall state on timelines, having meditated upon it, is basically "wanna bet?". I'm not sure what else I could say to drill into the crux of the issue, without knowing more where Joe and I disagree.

The gap between AI systems then and AI systems now seems pretty plausibly greater than the remaining gap

Joe Carlsmith: I'm sympathetic at an intuitive level, but I also have wide error bars at this level of abstraction. In particular, have in mind worlds where some task turns out to be "very hard," despite cool progress. When I try doing this sort of extrapolation with brain emulation, for example, I think it's pretty plausible that despite >50 years of pretty cool neuroscience progress, we're still less than halfway (at least in researcher-years) to emulating human brains (modulo some other automation of scientific progress more broadly), because it turns out the brain is pretty gnarly, difficult to understand, and just requires a large amount of grunt-work, data-gathering in the face of technical/logistical limitations, etc. And especially if I set aside specific views about ML being promising, it seems plausible to me that AI turns out to be like brain emulation in this respect.

See e.g. Ajeya's stuff on the limitations of "subjective impressiveness extrapolation":

https://docs.google.com/document/d/1cCjzZaj7ATbq8N2fvhmsDOUWdm7t3uSSXv6bD0E_GM/edit#heading=h.njuz93bimqty.

[gap between us and] AGI is made mostly of intangibles.

Joe Carlsmith: Though I think there are some selection effects, where we have incentives to develop better tangible benchmarks re: what AI systems can't do, as they become more capable and can do more stuff.

And in a basic sense, if you look out at the space of tasks humans actually perform in the world, there's a huge amount of stuff that AIs can't do. Here I'm expecting you'll lean on all of these tasks showing up when we're a hair's breadth away from full AI already, but it seems like this requires substantive background hypotheses about the trajectory of AI capability development, such that you expect human cognitive labor to mostly get automated "all at once" rather than in a gradually increasing way. And there I'd want to know what's underlying those hypotheses. For example, on Ajeya's "horizon lengths"-centric picture, everything doesn't come "all at once" -- rather, you get "short" tasks substantially before you get "longer" ones. And we can imagine further limitations on e.g. data, environments, compute, etc, even if in some sense we know how to do it "in principle".

it looks like experts pointing to specific technical hurdles

Joe Carlsmith: Though it feels like the ability to point to technical hurdles benefits significantly from understanding what you need to do well enough to point to specific hurdles that could get in your way, and it's not clear to me that we're there with A. E.g., with something like, say, very fast interstellar space-craft, we have enough understanding of the technology and basic dynamics to say stuff about required fuel, materials, etc. But with AI it feels like we're groping around to a degree that I wouldn't trust to recognize technical hurdles even if they were there. And when we do condition on a specific pathway we this will get us there -- e.g., "X method of scaling up ML systems will basically work" -- there are indeed specific candidate hurdles re: compute, data, environments, etc that emerge from rough extrapolation from current scaling laws and other assumptions (though it's definitely possible to debate exactly what to expect in these respects).

exited that regime a few years ago.

Joe Carlsmith: What are the technical hurdles you think we cleared a few years ago? Stuff like image recognition and Go? I would've thought that you could've made arguments like this for decades, and that re: image recognition and go, you would've just said "we don't know of any specific technical hurdles to automating image recognition and go, so who knows when it'll get done."

Human-brain level hardware is one candidate, but how much this is really a technical barrier has always been an open question (as has the relevant quantity). And from your comments below, it doesn't sound like it's your focus.

[Partial summary of Nate's reply: Nate affirms that he has hurdles like those in mind. More generally, he has in mind "stuff the human brain does easily in a half-second".]

but all the following things should have happened first, before I made that observation!".

Joe Carlsmith: Before e.g. image recognition and Go, would you have said "but image recognition and Go should have happened first"?

In that case, I start to wonder about how to tell the difference between image recognition/go, and the plethora of tasks we can currently do.

Overall though I'm sympathetic to "I just wouldn't be that shocked by 5 years" (for me these are centrally "scaling up GPT-3 type systems works" worlds). 2 is pushing it for me, especially for "can easily nano-tech you" systems.

[*Partial summary of Nate's reply:* Nate does think that "all this 'AI' field can do is brute-search look a bunch of ply deeper into game tree than a human; it can't even parse up a visual field into shapes and edges in a manner that allows it to practically use that information" is a reasonably solid "we aren't on the brink of AGI" argument. Nate doesn't feel that the remaining hurdles have the same caliber of "I was supposed to see that happen many years before I saw AGI".

Nate concedes that his statement was slightly exaggerated -- there are certain brands of theorem-proving such that Nate would be willing to stick his neck out and say "this comes just before the end".

In 2009, however, Nate reports that you could find people who would scoff and say "AI can't even X, it must be far". Now, if you press people for analogous hurdles, people's response is to try to weasel out of it, because they just don't know. Nate is trying to point to a different social atmosphere, where if you were at conferences before 2014 and after 2014, you would notice a shift from it being cool to boldly claim that true AI is obviously far off because of X, Y, and Z, to people nervously avoiding being pinned down on the topic of what will not be doable in a few years' time.

None of this, on Nate's view, means that AGI is definitely right around the corner. But Nate takes it to mean that he doesn't get be indignant if reality says "and all the rest of the stuff falls very quickly once you have one or two more key insights".]

there's a decent chance that insight and innovation have a big role to play

Joe Carlsmith: Though note that Ajeya's report includes an adjustment for algorithmic progress:

https://docs.google.com/document/d/1cCjjzAj7ATbq8N2fvhmsDOUWdm7t3uSSXv6bD0E_GM/edit#heading=h.epn531rebzyy

There has been a lot of AI progress recently. When I tried to adjust my beliefs so that I was *positively* surprised by AI progress just about as often as I was *negatively* surprised by AI progress, I ended up expecting a bunch of progress.

Joe Carlsmith: Broadly sympathetic to this.

"society might collapse

Joe Carlsmith: Fwiw, this isn't playing any role for me.

also things might speed up a bunch when we get close & can build tools that help us speed up

Joe Carlsmith: Interesting. As I mentioned above, something like this plays an important role in some models OP/our advisor are interested in, and it feels to me like an important question re: take-off dynamics more broadly. Not sure it's the best use of time re: this report in particular, but would be interesting to hear your take in more detail at some point.

[*Partial summary of Nate's reply:* On Nate's model, by the time AI is able to contribute significantly to AI progress, you're squarely in the endgame. This sort of thing maybe makes a big difference as to whether the endgame lasts for three months versus three years. From Nate's perspective, however, that sort of difference doesn't make nearly as much of a difference in our survival odds as how well we're prepared when we hit the endgame.]

My overall state on timelines, having meditated upon it, is basically "wanna bet?". I'm not sure what else I could say to drill into the crux of the issue, without knowing more where Joe and I disagree.

Joe Carlsmith: My guess is that debating timelines isn't the best use of our time for present purposes, as I didn't spend much time on them in the report, relative to arguments for doom conditional on timelines. That said, to the extent that you're at 85% on a pretty demanding conception of AGI (e.g., "a single system that can nano-tech all humans very easily, starting from not-especially-different-from-today levels of knowledge and tech") by 2070, I do feel open to betting. I'm hoping to work out my own take on timelines better, either in the process of revising the report or sometime after (in the current version I'm mostly leaning on work by other folks), though, so would prefer to wait to see where I end up after that.

Difficulty of alignment

Re (2), on it being comparatively difficult to build aligned systems, I suspect a bunch of our disagreement is in the "at least superficially attractive to deploy" clause. In particular, I basically expect the "endgame" to be a shitshow, full of events at least as crazy-seeming-to-us as:

- The AI system is drawing up plans to Kill All Humans. The decision is to retrain the system until it doesn't seem to be drawing up such plans. Our community cautions that this will not train out the behavior, but only its appearance in our shallow diagnostic tools. The plan goes forward anyway.
- The people fretting about the clear misalignment of the AI systems being developed, have limited social capital inside the organization, and are met with skepticism and resistance when they highlight a wide variety of flaws, and they can't do much but watch and protest as the team builds a system that is deeply and hopelessly misaligned.
- The AI system is clearly not safe for being deployed, but a state actor already has a copy of the code, and they've deployed it in a risky way, and will probably deploy it in an even more risky way next week, and the team decides that their deployment is overall less risky than that.

And more realistically, quite plausibly as crazy as "running the AGI on internet-connected computers, unfettered".

Like, on my model, the AI doesn't *need* to look all that superficially good. The relevant actors will feel forced into corners where the good-looking-ness is only tangentially relevant. The social, political, and psychological forces pushing people to persuade themselves that things look good despite the superficial evidence to the contrary will likely be large. The future is likely to look *derpy* rather than *competent*; I predict it will be more like a WWII story from Catch 22 than a WWII story from Saving Private Ryan (eg, a story where everyone is fumbling and bumbling, rather than a story in which everyone is dutifully and honorably carrying out a well-defined purpose).

I suspect we have some object-level disagreement on this point, though I'm not quite sure where. Perhaps Joe disagrees with the view that the modern world is more like Catch 22 than Saving Private Ryan? Perhaps he expects it to change in the endgame? My counters are roughly "behold the 2016 elections" and "behold the world response to the coronavirus". To get into more detail, I'd want some back-and-forth.

The AI system is drawing up plans to Kill All Humans. The decision is to retrain the system until it doesn't seem to be drawing up such plans. Our community cautions that this will not train out the behavior, but only its appearance in our shallow diagnostic tools. The plan goes forward anyway.

Joe Carlsmith: I think maybe we see stuff analogous to this; but I think we do have some different levels of pessimism here. That is, I think that if e.g. Demis/Sam literally observes an AI system drawing up plans that literally involve killing all the humans (e.g., "step 3: here all the nano-bots burst forth from the human bloodstreams"), I think there would be a substantial bigger freak-out than just "ok let's retrain it until it doesn't make plans like this." For less blatant forms of bad-intended-behavior, though, it's less clear. The type of race dynamic we're in also matters here.

The AI system is clearly not safe for being deployed, but a state actor already has a copy of the code, and they've deployed it in a risky way, and will probably deploy it in an even more risky way next week, and the team decides that their deployment is overall less risky than that.

And more realistically, quite plausibly as crazy as "running the AGI on internet-connected computers, unfettered".

Joe Carlsmith: This seems pretty imaginable.

Perhaps Joe disagrees with the view that the modern world is more like Catch 22 than Saving Private Ryan?

Joe Carlsmith: I think we probably do have some background disagreement here. E.g., I expect that you are more broadly pessimistic about the world's competence than I am, even though I don't think the world is a saving-private-ryan situation. Might be interesting to try to pin down the disagreement into specific predictions. The "what do people do if they literally see the AI system trying to kill people" case is maybe one example.

And note that some things one might class as "incompetence" can push in the good direction in the case of AI risk: e.g., people being naturally suspicious of new/scary/ethically weird technology (cf. nuclear power, cloning), very strong backlash to small incidents (e.g., nuclear power, accidents with self-driving cars), intense regulation, etc.

"behold the 2016 elections" and "behold the world response to the coronavirus"

Joe Carlsmith: Agree that these are instructive data points.

Another point of potential disagreement is that I don't think AI alignment needs to be all that hard to make proposition (2) true-enough-to-kill-us. Like, I'm not expecting alignment to be a feat of herculean difficulty unlike any technical obstacle humanity has ever faced in the past. I'm just expecting it to be *hard enough* that we can't pull it off between the time the first team gets a sufficiently capable AGI, and the time that someone else deploys one (or they do, either out of derpy incompetence or crappy game theoretic response to an unwinnable gameboard).

From my perspective, the claim "maybe it won't be very hard to align these things" sounds a lot like "maybe my code won't have bugs". It seems like sheer blind optimism, in the face of evidence like "if you train humans real hard for reproductive fitness they still invent condoms", to expect alignment to be so easy that it can be carried out last-minute, in the window between having-AGI and the end of the world.

Perhaps Joe expects alignment to be solved before that time window? (I don't; it hasn't been going very well so far, from my perspective. Perhaps we have a disagreement about how promising the research currently being done is, or how promising it's going to get before the endgame.)

Perhaps Joe expects that time window we'll have is significantly and relevantly longer than I expect? (I expect that would take a huge effort of civilization-wide coordination that our civilization seems categorically incapable of organizing.)

Perhaps Joe thinks the relevant time window will be long enough without a massive coordination effort, because he expects all relevant researchers will understand that they're handling a civilization-ending catastrophe device rather than a petulant system that will turn into a pot of gold if it gets shaken up a bit? (I don't; I expect the AI researcher culture to be about as competent around alignment issues as they are now, or near enough as makes no difference.)

Perhaps Joe thinks that alignment is so easy that it can be solved in a short time window?

My main guess, though, is that Joe is coming at things from a different angle altogether, and one that seems foreign to me.

Attempts to generate such angles along with my corresponding responses:

- Claim: perhaps it's just not that hard to train an AI system to be "good" in the human sense? Like, maybe it wouldn't have been that hard for natural selection to train humans to be fitness maximizers, if it had been watching for goal-divergence and constructing clever training environments?
- Counter: Maybe? But I expect these sorts of things to take time, and at least some mastery of the system's internals, and if you want them to be done so well that they actually work in practice even across the great Change-Of-Distribution to operating in the real world then you've got to do a whole lot of clever and probably time-intensive work.
- Claim: perhaps there's just a handful of relevant insights, and new ways of thinking about things, that render the problem easy?
- Counter: Seems like wishful thinking to me, though perhaps I could go point-by-point through hopeful-to-Joe-seeming candidates?

I dunno, I might be able to generate more by thinking hard about it, but it would be much easier to find someone who disagrees (perhaps Joe) and have a bit of a back-and-forth.

Joe expects that time window we'll have is significantly and relevantly longer than I expect?

Joe Carlsmith: Depending on "relevantly longer," I think there's some of this. By default I expect years of work with systems that are sufficiently APS that you're actually getting relevant empirical data about the real problem, learning important stuff, making progress, and so on.

That is, in my mainline model you don't need to wait to develop some super advanced system, then put things "on pause." Your empirical alignment work is getting incrementally more useful along the way (and you're getting incrementally more useful assistance from AI tools, weaker agents, etc.).

thinks the relevant time window will be long enough without a massive coordination effort, because he expects all relevant researchers will understand that they're handling a civilization-ending catastrophe device rather than a petulant system that will turn into a pot of gold if it gets shaken up a bit?

Joe Carlsmith: I do tend to expect significant delays between development of a given AI system, and its large-scale intention deployment in the real world, at least for many applications. For example, I don't think the military will just take the first advanced AI system it gets its hands on and put all the nukes in its hands. See also self-driving cars, delays in the adoption of electricity, etc. But I think this centrally due to expecting various frictions/adoption delays/regulatory hurdles/bureaucratic problems, plus fear-of-new-tech dynamics, rather than "everyone is super responsible and cautious re: alignment X-risk in particular."

Perhaps Joe thinks that alignment is so easy that it can be solved in a short time window

Joe Carlsmith: I do have some probability that the alignment ends up being pretty easy. For example, I have some probabilities on hypotheses of the form "maybe they just do what you train them to do," and "maybe if you just don't train them to kill you they won't kill you." E.g., in these worlds, non-myopic consequentialist inner misalignment doesn't tend to crop up by default and it's not that hard to find training objectives that disincentivize problematically power-seeking forms of planning/cognition practice, even if they're imperfect proxies for human values in other ways.

My main guess, though, is that Joe is coming at things from a different angle altogether, and one that seems foreign to me.

Joe Carlsmith: My main guess is more like: I place more weight than you on comparatively optimistic setting for variety of different variables -- e.g., alignment easiness, timelines, take-off speed/discontinuity, possibility of non-APS systems being th

main thing, ease of DSA/nano-teching, civilizational competence (or over-caution-type incompetence that ends up being helpful), correction ability, some crucial aspect of this discourse being confused/mistaken -- such that I end up with significant credence on "no" for lots of premises where you're at >90%, and this adds up.

perhaps it's just not that hard to train an AI system to be "good" in the human sense? Like, maybe it wouldn't have been that hard for natural selection to train humans to be fitness maximizers, if it had been watching for goal-divergence and constructing clever training environments?

Joe Carlsmith: I think something like this is in the mix for me. That is, I don't see the evolution example as especially strong evidence for how hard inner alignment is conditional on actually and intelligently trying to avoid inner misalignment (especially in its scariest forms).

Change-Of-Distribution to operating in the real world

Joe Carlsmith: One question I have here is whether "operating in the real world" really ends up as a binary/irrevocable switch. That is, it seems like all along the way, we are probably going to be getting data about what it's like to deploy systems in the real world, learning about what goes wrong, deploying them in limited/controlled contexts, revoking their spheres of influence when we see problems, etc.

Of course, for any given choice to develop an AI system or extend its influence, including e.g. letting beta-users access the GPT-3 API, there is some probability that "that step was the fuck-up and now everyone is going to die." But the question of what we should actually expect that probability to be at each actual stage, given the work done and understanding gained by that point, seems like a further question -- and I worry that imagining a "big red deploy button" that we press and then "cross our fingers" because we basically have no data about how this is going to go" will end up a misleading image.

perhaps there's just a handful of relevant insights, and new ways of thinking about things, that render the problem easy?

Joe Carlsmith: This isn't a big part of it for me. I don't expect some conceptual "click" that solves the problem.

Misalignment outcomes

Re (3), on whether deployed misaligned systems are catastrophic, it looks to me like I have two points of disagreement with Joe:

1. I suspect I think that the capability band "do a trillion dollars worth of damage, but don't Kill All Humans" is narrower / harder to hit.

2. I suspect we disagree about how much warning shots help civilization get its act together and do better next time.

With respect to the first point, this might be because I'm somewhat more bullish about rapid capability gain, or it might be due to some of the points listed above about the usefulness of cognition changing sharply as the quality of cognition increases slowly (shitty analogy: if "1 million 6-year old hours" is still useless, but "1 million 14-year old hours" is not, and we spend all our time going from rat-level capabilities in some domain to monkey-level capabilities in that domain, we might blow straight past the relevant threshold in a weekend). As such, on my picture, trillion-dollar warning shots just don't happen all that often, and so even if society *would* get its act together in the face of one, we won't have the opportunity.

(Or, well, a trillion dollars of damage is not all that huge a number these days, but whatever, we can work with 10 trillion instead.)

1. I suspect I think that the capability band "do a trillion dollars worth of damage, but don't Kill All Humans" is narrower / harder to hit.

Joe Carlsmith: I feel sympathetic to points in this vein, and have been thinking for a bit about revising my probability on premise 5 to reflect greater correlation between "trillion dollars of damage worlds" and "full disempowerment worlds."

And of course I also disagree that society *would* get its act together in the face of warning shots. As case-in-point, I exhibit the global response to the coronavirus: it was a 10 trillion dollar warning shot about pandemics. Is society now going to get its act together with regards to biological risks? Is gain-of-function research going to be stopped, conclusively, world-wide? Is machinery for rapid development and deployment of vaccines going to be built and maintained? Get your bets in now!

Because my bet is: lol no. Not even close. And if civilization can't ban gain-of-function research, when it has only very dubious benefits and huge risks, and when the forces arrayed in favor are a mere handful of academics, then why should I expect that civilization will coordinate around making sure that AI research is safe, when AI technology is having large and visceral positive

impacts on society (or at least the economy), and the forces arrayed in favor are enormous business interests with huge amounts of money to gain?

In short, if we are lucky enough for our first AI failure to be a mere warning shot rather than a civilization-ending catastrophe, I expect our civilization to do with it the same thing they do with every other warning shot: squander it completely. It probably won't even make a dent in the institutional inertia behind pushing research forward, and even if it did then the business interests arrayed behind AI research proceeding, and the psychological forces pushing researchers to believe in their research, and the sheer ease of saying things like "well that system only made that mistake because it was too stupid", will be more than enough to overpower any resistance. Or so I predict.

And if civilization can't ban gain-of-function research, when it has only very dubious benefits and huge risks, and when the forces arrayed in favor are a mere handful of academics, then why should I expect that civilization will coordinate around making sure that AI research is safe, when AI technology is having large and visceral positive impacts on society (or at least the economy), and the forces arrayed in favor are enormous business interests with huge amounts of money to gain?

Joe Carlsmith: I find this example fairly compelling.

every other warning shot

Joe Carlsmith: Civilizational reactions to much smaller nuclear disasters like Chernobyl and Three Mile Island seem like an instructive data point here.

It probably won't even make a dent in the institutional inertia behind pushing research forward, and even if it did then the business interests arrayed behind AI research proceeding, and the psychological forces pushing researchers to believe in their research, and the sheer ease of saying things like "well that system only made that mistake because it was too stupid", will be more than enough to overpower any resistance.

Joe Carlsmith: I think our intuitions here are different. I think that if a rogue AI system, for example, crashed the financial system and tried to bioweapon everyone, but only killed 50M people instead of everyone, and then was finally gotten under control via some extreme civilizational effort like turning off the entire internet and destroying tons of computers, and this was known and vivid to the world, this would put a very significant dent in the institutional/research inertia and business interest pushing for just continuing to forward in scaling up similar systems. And I really don't expect "that just happened because it wasn't smart enough let's just make it smarter."

Indeed, this is the kind of thing I can readily imagine leading to really extreme worldwide backlash, intense international coordination, research bans, nations threatening to bomb other nations if they build sufficiently big compute clusters, etc.

[*Partial summary of Nate's reply:* Nate agrees that this particular scenario would perhaps disrupt the narrative enough, but he predicts that this won't happen because it's too narrow a capabilities target to hit.]

Now, question (3) might be a place where I'm getting a little bitten by my substitution of Joe's "APS systems" and my "Scary systems" (in which case more of our disagreement is shunted into question (2), b/c not only do I have a higher probability of AI systems being sufficiently-superficially-aligned that some fool is deploying them, but I also believe the deployed systems are drawn from a scarier class), but I also get the sense that we have a disagreement about general civilizational competence, and its ability to react in sane and reasonable-seeming ways when the stakes are high and a bunch of value is on the line.

As usual, I could dig deeper into various points where I suspect disagreement, but I'm not sure where the real disagreements lie, so for now I'll desist.

Argument style

Now, critiquing the argument style: I worry that Joe's style of argumentation, of breaking a proposition into a series of (allegedly conditional) conjunctive steps and then multiplying to attain an estimate, has a tendency to give answers that are much too low. (Joe acknowledges this point himself in the report, though does not combat it to my satisfaction. This is known as the "multi-stage fallacy", and its namesake -- and an example of its use in the wild -- is exhibited by Nate Silver here: <https://fivethirtyeight.com/features/donald-trumps-six-stages-of-doom/>.)

Speaking roughly, if we break an argument into n conjunctive steps, and try to look "reasonably uncertain" about each step, this will tend to drive our probability of any event happening to around 1×2^{-n} . In particular, with a six-step breakdown (as in the main document) the "reasonable uncertainty" drives the answer towards $1 \text{ in } 64$ or about 1.5%, and a three-step breakdown (as in the appendix) drives the answer towards $1 \text{ in } 8$ or about 12.5%. 5% is comfortably in the middle of those two ranges (and strikingly close to their geometric mean), and I worry that the probabilities assigned in the report are mostly an exercise in deploying the multi-stage fallacy.

One intuition for combating this fallacy is that we're supposed to make heavier use of disjunction as well as conjunction in our models. Another is that, when conditioning on earlier stages, our probabilities are supposed to get *so extreme* that our overall probability could not be driven down further by someone decomposing the claim into further conjunctions that apparently partition the space.

One intuitive motivation for the latter is that the realities that managed to get one or two surprises deep into our list of stages, likely contain some underlying force driving a correlation between all the stages, that spikes the conditional probabilities in the later stages. (In particular, I diagnose Nate Silver's error in the post linked above.)

A further intuition-pump here is that the future has a good chance of surprising us or going sideways, and if it instead falls neatly into the first one or two stages we named clearly in advance, then whatever process picked out the whole series of stages was probably onto something, and *conditional* on the first few stages, "model uncertainty" (and other forces driving our error bars to be "virtuously wide") is much lower in the later stages.

(Indeed, my guess is that I myself have failed to account entirely for this phenomenon, as evidenced by a guess that my quick-and-dirty numbers on Joe's six-stage argument would yield a lower implied probability than 77%. Let's check! Eyeballing the six stages and choosing numbers off the cuff to 5% precision, I assign... 85%, 100%, 90%, 95%, 95%, 100%, for a total of around 69% -- a decent difference. As I said before, my actual probability of catastrophe is larger, both from leakage to disjuncts technically excluded by Joe's breakdown, and due to a sense that my conditional probabilities would likely drift higher on reflection as I account for the multi-stage fallacy.)

It seems to me that Joe perhaps attempted to combat the multi-stage fallacy (perhaps due to critiques given by my colleagues and me, on a draft version of this report). In particular, in his appendix, Joe not only makes a 3-stage rather than 6-stage argument, but also considers a "positively-phrased" argument (in both a 3 and 6 stage breakdown).

The three-stage argument does a bit to address my concerns -- it is better to use an argument style that drives all probabilities to 12.5% than 1.5%. The "positive framing", however, does not at all address my critique. In particular, Joe's positively-phrased argument is not conjunctive, but rather disjunctive!

Just as a multi-stage conjunctive argument of length n drives all "reasonably uncertain" assignments of probabilities towards 1 in 2^n , a multi-stage *disjunctive* argument of length n drives all "reasonably-uncertain" assignments of probabilities towards $(2^n - 1)$ in 2^n , ie 63/64 for 6-stage arguments and 7/8 for 3-stage arguments! So it is no consolation to me that his conjunctive 5% turns into a disjunctive 95%.

(Of course, perhaps Joe was merely attempting to combat certain framing effects, and not also the multi-stage fallacy, in which case my critique still stands, but critiques the argument proper rather than some flawed response to prior criticism.)

An example of a conjunctive, positively framed argument might be:

For humanity to make it to 2070 alive, we need all three of:

1. at least 20 years to prepare, *and*
2. the technical challenge of alignment to be pretty easy, *and*
3. the research culture to be alignment-conscious in a competent way.

Someone attempting to be virtuously uncertain might assign probabilities like, say, 50%, 75%, and 60%, implying a mere 22.5% chance of survival. Now, I'm not saying those are my numbers (they aren't); I'm saying that this is what a "positively framed" conjunctive argument feels like. And, from my perspective, the way one counteracts the multi-stage fallacy is not to simply shorten their conjunctive breakdown or explicate its disjunctive counterpart, but rather to consider also conjunctive breakdowns of the counterpoint. The difficulty is not in reconciling one's conjunctive account of catastrophe with their disjunctive account of survival, but in reconciling a conjunctive account of catastrophe with a conjunctive account of survival.

The fact that Joe frames his catastrophe estimates conjunctively, and his survival estimates disjunctively, does little to assuage my fears that the final probability in his report is driven ultimately by his choice of which side gets the conjunctions. Joe admits in his report that these are lower bounds, but seems to feel they are not too far off. By contrast, I fear that his style of argumentation has driven them quite a bit too low.

and I worry that the probabilities assigned in the report are mostly an exercise in deploying the multi-stage fallacy.

Joe Carlsmith: A bit about my backdrop model here, which I think may differ from yours. On my general model, arguments have a kind of "true conjunctiveness/distinctiveness," which it is the task of argument formulation to capture. Thus, for example, the argument that "next new years day you will get hit by lightning while getting eaten by a shark, shortly after winning the lottery" is genuinely conjunctive: you really should be breaking this into conjuncts and estimating their probability-independently. And there are other arguments -- "here is my overly-specific forecast about how Y technology-i-have-no-control-over will get developed in the next 5 years" -- that are more subtly like this.

Thus, it's not enough to say "thou shalt not break an argument into many stages" or "thou shalt focus on disjunctive formulations rather than conjunctive formulations." Rather, there is a kind of discriminating taste involved in knowing how to carve an argument "at the joints," in a way that brings out the amount of conjunctive-ness it actually involves, but not more, and which makes it easy to capture the correlations at stake in the premises (including the correlations implied by the fact that the person making the argument is offering it at all; e.g., if a CEO tells you an overly-specific plan that they will use to get the company to succeed, there are indeed correlations at stake re: their being able to push for the plan, their being able to identify it ahead of time, etc.).

To the extent that there are underlying factors driving correlations between the premises, a good formulation brings those out and makes them into premises in themselves. Thus, for example, if "Trump is actually a really effective campaigner" is the thing that ultimately drives his probability of getting the nomination, we should bring that out in particular and ask what probability we place on it (though you do also want to capture the worlds where Trump gets the nomination without being an effective campaigner -- so you need some disjunction in that sense).

(And it's true that in the lightning/lottery example above, if you do in fact win the lottery shortly before new years day, you should update hard in favor of whoever made this argument having some sort of spooky foresight/power or whatever, thereby introducing new correlations.)

[Partial summary of Nate's reply: Nate agrees that "some arguments have the conjunctive nature, and some arguments have disjunctive nature". He says that his point is that he thinks Joe's numbers are far too low, and thinks this is related to Joe using disjunctive breakdown for a "we will survive" argument that is more properly conjunctive. Nate claims that using a disjunctive breakdown will tend to push Joe's survival probabilities way higher, barring significant skill in extremizing conditional probability that Nate does not see evidenced.]

disjunction as well as conjunction in our models

Joe Carlsmith: Fwiw, I don't put a lot of weight on the idea that my premises actively miss a lot of disjunctive routes to power-seeking X-catastrophe, such that we should have a lot of probability mass on power-seeking X-catastrophe coming from world where one of my premises is false. In particular, it feels to me like my timelines condition is fairly minimal (and close to required for the instrumental convergence argument to go through); like the "trillion dollars of damage", "full scale disempowerment," and "disempowerment = catastrophe" are implied by basically any X-catastrophe story, and that basically all mainline doom stories involve significant incentives to build the relevant systems, and hard alignment problems.

My main candidates for scenarios that "slip through the cracks" are ones where such incentives aren't there to the right degree and/or where alignment is easy but we mess up anyway -- but I don't feel like these scenarios are driving close to >50% of the probability mass on power-seeking doom.

That said, I do think it's worth playing around with versions of that argument that focus less on things that strictly need to happen, and more on candidate factors that could drive correlations between premises (e.g., timelines, civilizational competence, take-off, etc), but where there is still significant p(doom) even if they're false (and so the argument ends up disjunctive in that sense). I'm hoping to do more of this going forward.

A further intuition-pump here is that the future has a good chance of surprising us or going sideways, and if it instead falls neatly into the first one or two stages we named clearly in advance, then whatever process picked out the whole series of stages was probably onto something, and *conditional* on the first few stages, "model uncertainty" (and other forces driving our error bars to be "virtuously wide") is much lower in the later stages.

Joe Carlsmith: I agree with this, and would like to think more about what might play this sort of role. The main salient candidates in my mind are something like: "Eliezer is right about stuff in general," "fast/concentrated take-off," and maybe "general civilizational incompetence."

Would be curious if you think there are others worth highlighting.

That said, variables like "absolute alignment easiness" seem pretty uncorrelated, in a metaphysical sense, with variables like "timelines" and "quality of civilizational response to the problem." So my best candidate correlation mechanisms there are more epistemically-flavored: e.g., "Eliezer is right about stuff."

The three-stage argument does a bit to address my concerns -- it is better to use an argument style that drives all probabilities to 12.5% than 1.5%.

Joe Carlsmith: Fwiw, to me it feels like the intuitive argument has at least two separate stages in its deep structure: e.g., AI needs to happen, and we need to all die as a result. And I'm inclined to think three is pretty minimal as well. That is, it feels to me like something like "AI will happen by 2070," "Alignment is a problem and non-trivial to solve" and "We will fail and all die" are pretty importantly separate claims that the argument basically requires, and that reasonable people can have significantly uncorrelated uncertainties about (I also feel this way about the 6 premise argument). And I feel like three premises is sufficiently short that the charge of "artificially" extending the argument so as to drive the probability lower rings false, at least to my ear.

perhaps Joe was merely attempting to combat certain framing effects, and not also the multi-stage fallacy

Joe Carlsmith: Indeed; the positively framed version was centrally meant to address framing effects where people will think it's more virtuous to be "skeptical" in some sense, and so to put low probabilities on claims, whatever they are.

An example of a conjunctive, positively framed argument might be:

Joe Carlsmith: Thanks for offering this, I found it helpful.

Would also be interested in any other formulations you have up your sleeve, when you're trying to boil down the basic case.

For humanity to make it to 2070 alive, we need all three of:

Joe Carlsmith: I think this version of the argument is going to be driven centrally by the probability one assigns to the overall claim that in order for humanity to survive, one needs all of these things -- probability which pretty clearly shouldn't be 100%

For example, there is surely some probability of survival if timelines are within 20 years; some probability that a not-that-competent research community could solve a "pretty easy" version of the alignment problem; and so on.

I'd like to think more about what a version of this argument I'd endorse would look like, and what probabilities I'd assign. One nitpicky issue is that it's not immediately clear to me what "probability you need all three of these things" actually means. E.g. is it something like "probability that absent these three things, my probability on doom should be ~100%"? But given that I'm assigning some substantive probability to worlds where all three things aren't true, it seems unclear what sort of probability is at stake in the "should" above (in general, I don't like assigning subjective probabilities to subjective probabilities -- e.g., "I'm 50% that my probability is >80% that p"). That said, could maybe formulate in terms of "if I thought about it X amount more, get to >Y% on 'you need all three of these things, else doom.'"

Joe Carlsmith: An alternative argument:

- (1) AGI by 2070
- (2) Eliezer and others at MIRI think >X% doom, conditional on AGI by 2070. What's your probability that they're right?

And, from my perspective, the way one counteracts the multi-stage fallacy is not to simply shorten their conjunctive breakdown or explicate its disjunctive counterpart, but rather to consider also conjunctive breakdowns of the counterpoint. The difficulty is not in reconciling one's conjunctive account of catastrophe with their disjunctive account of survival, but in reconciling a conjunctive account of catastrophe with a conjunctive account of survival.

Joe Carlsmith: I found this general framing helpful, thanks. That said, as mentioned above, I do feel like the right approach to this type of dynamic is specific-argument-dependent: e.g., some arguments just are more conjunctive in one formulation vs. another (see the "hit by lightning while winning the lottery" argument above).

In closing, I again note that I found Joe's report to be remarkably well-reasoned and thorough. I have harped on a variety of points of disagreement, but let us not lose sight of how much we do agree on, such as the overall framing of the problem, and what sorts of questions to be asking and arguments to be making. I see the analysis he uses to support his conclusions as a good breed of analysis; it has a form that takes into account many of the features of the strategic landscape that seem salient to me.

Joe Carlsmith: Glad to hear it, and thanks for your comments!

EfficientZero: human ALE sample-efficiency w/MuZero+self-supervised

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

This is a linkpost for <https://arxiv.org/abs/2111.00210>

["Mastering Atari Games with Limited Data"](#), Ye et al 2021:

Reinforcement learning has achieved great success in many applications. However, sample efficiency remains a key challenge, with prominent methods requiring millions (or even billions) of environment steps to train. Recently, there has been significant progress in sample efficient image-based RL algorithms; however, consistent human-level performance on the Atari game benchmark remains an elusive goal.

We propose a sample efficient model-based visual RL algorithm built on [MuZero](#), which we name **EfficientZero**. Our method achieves 190.4% mean human performance and 116.0% median performance on the Atari 100k benchmark with only two hours of real-time game experience and outperforms the state SAC in some tasks on the DMControl 100k benchmark. This is the first time an algorithm achieves super-human performance on Atari games with such little data. EfficientZero's performance is also close to DQN's performance at 200 million frames while we consume 500 times less data.

EfficientZero's low sample complexity and high performance can bring RL closer to real-world applicability. We implement our algorithm in an easy-to-understand manner and it is available at [this https URL](#). We hope it will accelerate the research of MCTS-based RL algorithms in the wider community.

This work is supported by the Ministry of Science and Technology of the People's Republic of China, the 2030 Innovation Megaprojects "Program on New Generation Artificial Intelligence" (Grant No. 2021AAA0150000).

Some have said that poor sample-efficiency on ALE has been a reason to downplay DRL progress or implications. The primary boost in EfficientZero ([table 3](#)), pushing it past the human benchmark, is some simple self-supervised learning ([SimSiam](#) on predicted vs actual observations).

Omicron Variant Post #2

It's now been three days since [Post #1](#). The situation is evolving rapidly, so it's time to check in. What have we learned since then? How should we update our beliefs and world models? There will inevitably be mistakes as we move quickly under uncertainty, but that's no reason not to do the best we can.

Update Update

What should we look for here and in the coming days?

1. No news is good news.Omicron is scary because many scary things are possible. The worse things are going to get, the sooner they will make themselves known. When we get news that something has happened, especially news that isn't the result of a lab or statistical analysis, that will typically be bad news, but we *expect* a certain rate of such bad news. If we get less than expected, that's good news.
2. The pattern of where and how we find cases, and the details of those cases, will give us better insight into how widely and quickly, and in which directions, Omicron has spread. This will tell us how far along things are, and give a better estimate and narrower range of how infectious Omicron can realistically be.
3. Information on how deadly Omicron is should update us quickly, and matters a ton for how to react, but beware of confounding factors, motivated statements and misunderstandings in both directions, and small sample sizes. Hospitalizations are not a natural category and at this stage deaths will be rare unless things are much worse than we expect. How mild the cases we find are will largely depend on how we are testing. Details matter.
4. The reaction of various countries, the stock market and others will continue to tell us what they think is going on, what they expect to happen next, and how they respond when such things happen. The general vibe of elite and media messaging also reveals much information, with a focus on what kinds of actions they will try to engineer rather than the 'facts on the ground.'
5. Information on immune erosion will come from results from laboratory analysis and other looks at Omicron's mutations, and other physical detail information, such as testing antibodies, and from statistics on who gets infected or sick once we have enough cases for that to mean something. It also comes from knowing how infectious Omicron is, since if we know how fast it's spreading and we know how fast it spreads to the immunologically naive, we also then know how immune erosive it is because math. The vaccine reactions will also tell the story - we should expect manufacturers to react to the situation based on what they know, and to some extent governments as well.
6. A lot of updating is having time to think, and integrate others who also have time to think. There doesn't need to be new information for our estimates to improve.

It's been announced that new versions of the vaccines are indeed coming, as they no longer expect the old ones to be fully effective, and are talking about a few months to get the new versions online.

Noah Smith gives his overview of the situation [here](#). It provides a lot of the same links, some good new links, a set of views broadly similar to mine, and is generally reasonable. I do disagree with the conclusion, in that I think on the margin the most valuable focus is now testing and treatment, with further vaccinations vital but not primary. Ramping up vaccine production capacity is still a fantastic idea. I will also be checking out his [list of Covid experts](#) on Twitter. A lot of them are not people I am already following.

[Here is Weekend Editor](#), author of many helpful comments and overviews, offering their best take, broadly compatible with mine, [including an addendum right at press time](#). An especially

worthwhile note there [via Your Local Epidemiologist](#):

Boosters not only raise the *number* of antibodies, but also the *diversity* of antibodies to various spots on the spike protein. [11] So it's important to get a booster once you're eligible: you not only get short- to intermediate-term increased immunity from antibodies, you also get long-term increased immunity because your antibodies check more carefully for mutations in the spike protein, giving broader immunity to variants. That's *in addition* to antibody maturation, where your immune system refines the antibodies over time to get better and better; your memory B-cells are *busy*.

[This set of slides](#) is mostly a good summary with some exceptions, such as (in the mind of my instinct viewer) incorrectly associating the current winter wave of Delta with the emergence of Omicron, when they are two completely distinct things.

Immune Erosion and Transmissibility

[Here is the WHO announcement on what we know](#), in which they reiterate all the things they would have said last week but deny that it's possible to know things under uncertainty:

Current knowledge about Omicron

Researchers in South Africa and around the world are conducting studies to better understand many aspects of Omicron and will continue to share the findings of these studies as they become available.

Transmissibility: It is not yet clear whether Omicron is more transmissible (e.g., more easily spread from person to person) compared to other variants, including Delta. The number of people testing positive has risen in areas of South Africa affected by this variant, but epidemiologic studies are underway to understand if it is because of Omicron or other factors.

Severity of disease: It is not yet clear whether infection with Omicron causes more severe disease compared to infections with other variants, including Delta. Preliminary data suggests that there are increasing rates of hospitalization in South Africa, but this may be due to increasing overall numbers of people becoming infected, rather than a result of specific infection with Omicron. There is currently no information to suggest that symptoms associated with Omicron are different from those from other variants. Initial reported infections were among university students—young individuals who tend to have more mild disease—but understanding the level of severity of the Omicron variant will take days to several weeks. All variants of COVID-19, including the Delta variant that is dominant worldwide, can cause severe disease or death, in particular for the most vulnerable people, and thus prevention is always key.

Effectiveness of prior SARS-CoV-2 infection

Preliminary evidence suggests there may be an increased risk of reinfection with Omicron (i.e., people who have previously had COVID-19 could become reinfected more easily with Omicron), as compared to other variants of concern, but information is limited. More information on this will become available in the coming days and weeks.

Effectiveness of vaccines: WHO is working with technical partners to understand the potential impact of this variant on our existing countermeasures, including vaccines. Vaccines remain critical to reducing severe disease and death, including against the dominant circulating virus, Delta. Current vaccines remain effective against severe disease and death.

An article on [Stat News](#) is a reasonable summary, and adds one key bit of information I hadn't seen at the time:

The WHO said Saturday that early evidence indicated the variant was causing reinfections at higher rates than other variants in South Africa, suggesting some ability to get around the immune response.

As a reminder, when the majority of people in your country are not vaccinated, this kind of observation very much isn't news:

Joe Phaahla, South Africa's health minister, said Friday that breakthrough infections were occurring in the country, but the majority of hospital admissions remained among people who were not vaccinated, suggesting that vaccines were still maintaining some level of protection against Omicron. But, he acknowledged, "it's still early days in terms of this particular variant."

It's possible that this is uncharitable and he was making a more meaningful observation, but he's not claiming to have done so here.

[Science offers a similar article](#), of the 'this is scary but we don't know anything for sure yet' variety, emphasizing that the rate of growth combined with the mutations are the reason we are so concerned at this early stage, but that the growth rate could mostly be a coincidence.

[Here's a similar thread of a lot of 'we don't know' strung together.](#) It's good to caution against overconfidence, but also important not to respond to uncertainty with only 'we don't know.'

[Here is a thread from Christian Althaus](#) asking what we can say about Omicron's potential advantages.



Christian Althaus @C_Althaus · Nov 28

...

The mutational profile of Omicron suggests a potentially significant transmission advantage. But can we already say something about increased transmissibility or immune evasion? I'll give it a try. Warning: Preliminary and based on VERY limited data. 1/15

B.1.1.529 – potential impact of mutations

-
- Multiple RBD and NTD mutations associated with **resistance to neutralizing antibodies** (and therapeutic monoclonal antibodies)
 - Cluster of mutations (H655Y + N679K + P681H) adjacent to S1/S2 furin cleavage site – associated with **more efficient cell entry → enhanced transmissibility**
 - nsp6 deletion (Δ 105-107) – similar to deletion in Alpha, Beta, Gamma, Lambda – may be associated with **evasion of innate immunity (interferon antagonism) → could also enhance transmissibility**
 - R203K+G204R mutations in nucleocapsid - seen in Alpha, Gamma, Lambda – associated with **increased infectivity**



Christian Althaus ✅ @C_Althaus · Nov 28

...

Replies to [@C_Althaus](#)

The observed rapid replacement of Delta by Omicron in the province of Gauteng in South Africa is suggestive of a transmission advantage. 2/15



Christian Althaus ✅ @C_Althaus · Nov 28

...

Fitting a multinomial logistic regression model to the proportion of different variants in South Africa results in an estimated growth advantage of Omicron of 0.43 (95% CI: 0.15-0.72) per day compared to Delta. 3/15



Christian Althaus ✅ @C_Althaus · Nov 28

...

[@TWenseleers](#) obtained a similar estimate of 0.38 per day. There are lots of caveats: Targeted sequencing, stochastic effects in low incidence settings, and successive superspreading events could significantly bias these estimates. But let's continue from here. 4/15



Christian Althaus ✅ @C_Althaus · Nov 28

...

Assuming the same generation time, the transmission advantage could act at two levels: 1) increase in transmissibility, 2) immune evasion. We recently developed a mathematical framework to relate differences in growth rates to these properties. 5/15

This is 0.38 to 0.43 advantage *per day*. That's a gigantic transmission advantage if anything like it holds up. Definitely well over 100% advantage the way it's typically measured.



Christian Althaus ✅ @C_Althaus · Nov 28

A potential increase in transmissibility can be expressed as $\rho D/R_w$, where X is the estimated growth advantage (0.43 per day), D the generation time (5.2 days), and R_w the effective reproduction of the previous variant (~ 0.8 in RSA during October). 6/15 ibz-shiny.ethz.ch/covid-19-re-in...

4

11

114



Christian Althaus ✅ @C_Althaus · Nov 28

This would result in an increased transmissibility of 280% (95% CI: 98-468%). With Delta having an $R_0 = 5-6$ in the Northern Hemisphere during winter, the R_0 of Omicron would be around 10-30. Not impossible, but such a jump seems rather unlikely. 7/15

5

56

223



Christian Althaus ✅ @C_Althaus · Nov 28

If the transmission advantage acted via immune evasion only, the level of immune evasion would be $\rho D(1-\Omega)/(\Omega R_w)$, with Ω being the proportion of the population that has fully protective immunity against infection with earlier variants. 8/15

1

13

137



Christian Althaus ✅ @C_Althaus · Nov 28

South Africa records an excess mortality of 230k during the pandemic, which corresponds to 0.39% of the overall population (github.com/dkobak/excess-...). We earlier estimated the infection fatality ratio for the population of South Africa to be 0.35% (doi.org/10.1007/s10654...). 9/15



Christian Althaus ✅ @C_Althaus · Nov 28

Hence, it is likely that almost everyone in South Africa has been infected with #SARSCoV2 and developed partial immunity against reinfection. In addition, 24% of the population have been fully vaccinated. 10/15

I am skeptical of 'almost everyone.' I can believe an IFR of 0.35% on its own, but some numbers: 89k confirmed deaths in South Africa out of 2.95mm confirmed cases, overall population 59.31mm. If almost everyone is infected then we're talking about missing >90% of all cases, and about half of all deaths, and the IFR remaining 0.35%, excluding reinfections, at that level of care. Still, I am willing to buy that we're missing a large majority of infections.

The WHO is saying that it looks like Omicron is 'causing reinfections at higher rates' but if almost everyone has already been infected in South Africa, then that's not the observation it

sounds like it is. It's instead a statement that those *previously diagnosed* are getting infected more often than those who previously had milder cases that went undiagnosed, which is very different in its implications. It also seems like evidence that South Africa *hasn't* mostly already been infected.



Christian Althaus @C_Althaus · Nov 28

...

Thus, the proportion of the population that is fully protected ('immune') against infection and further transmission must be quite high. If we assume $\Omega = 75\%$, we get an immune evasion of 93% (95% CI: 32-100%), i.e., Omicron evades protective immunity in 93% of individuals. 11/15

13

68

227



Christian Althaus @C_Althaus · Nov 28

...

For $\Omega = 90\%$, we obtain an immune evasion of 31% (95% CI: 11-52%) 'only'. This clearly illustrates the current level of uncertainty about Omicron, and I want to emphasize again the preliminary character of these calculations. 12/15

4

40

187



Christian Althaus @C_Althaus · Nov 28

...

Still, I do expect partial immune evasion to be the main driver of the observed dynamics, but increased transmissibility cannot be ruled out so far. 13/15

8

31

207



Christian Althaus @C_Althaus · Nov 28

...

The developments in South Africa and observations from other countries during the coming days and weeks will allow us to shed more light on the properties of Omicron. 14/15

3

15

143



Christian Althaus @C_Althaus · Nov 28

...

Finally, thanks for the amazing work from [@Tuliodna](#), [@houzhou](#), [@rjlessells](#), and their colleagues in South Africa, [@firefox66](#) from [@ISPMBern](#) at [@unibern](#) for [covariants.org](#), [@richardneher](#) and [@trvrb](#) for [@nextstrain](#), and the many others working around the clock. 15/15

I like how this points out the relationships involved and how different numbers need to relate to each other. Those relations don't change that much based on the share that had immunity, so long as that share was 'large.' It does emphasize that there's too many free variables, and for now we can't draw conclusions off an analysis like this, only constrain the set of possible worlds.

Several people pointed me to [this Times of India report](#):

JOHANNESBURG: The new [Omicron](#) variant of the coronavirus results in mild disease, without prominent syndromes, Angelique Coetzee, the chairwoman of the [South African Medical Association](#), told [Sputnik](#) on Saturday.

The [World Health Organization](#) (WHO) identified on Friday the new South African strain as one of concern, as it is reported to carry a high number of mutations -- 32 -- which possibly makes it more transmissible and dangerous. The WHO has dubbed it Omicron, the 15th letter of the Greek alphabet.

"It presents mild disease with symptoms being sore muscles and tiredness for a day or two not feeling well. So far, we have detected that those infected do not suffer loss of taste or smell. They might have a slight cough. There are no prominent symptoms. Of those infected some are currently being treated at home," Coetzee said

That would be the best case scenario. If Omicron only results in mild infections, we should welcome our new Omicron overlords as opposed to imposing new restrictions. However, it's way too early for that, and the cohorts involved seem like they were younger and healthier. [Here's what I believe is another take on the same statement, from Haaratz](#):

According to the report, omicron patients tended to be younger, and the variant was not found as often among the older population. Those infected with the variant mainly experienced fatigue and body aches. Still, it is still unknown what effect infection has on the older adults with underlying medical conditions such as diabetes or heart disease.

Prof. Dror Mevorach, head of the coronavirus department at Hadassah University Hospital Ein Karem, said the preliminary reports on the clinical condition of people infected with the new variant are encouraging. "If it continues this way, this might be a relatively mild illness compared to the delta variant, and paradoxically, if it takes over, it will lead to lower infection rates," and it will be easier to deal with globally.

[Here's another report, this time from the Telegraph](#). Once again, cohort young and healthy, but symptoms more mild than would have been expected given that.

[Here's a thread emphasizing that we can presume vaccines will continue to work against severe disease in Omicron](#).



Chise MFF @sailorrooscout · Nov 26

...

Why are variants unlikely to FULLY evade vaccine-induced immunity?

- Vaccines are POLYCLONAL
- CD8+ T-cells covering 52 epitopes across the spike protein
- CD4+ T-cells covering 23 epitopes across the spike protein

For more on this see: [science.org/doi/10.1126/sc...](https://science.org/doi/10.1126/science.abb3333)

[Goldman Sacks endorses that the vaccines will keep working, does not recommend portfolio changes](#).

Travel Restrictions

Inevitably we are seeing [a lot of this](#):



Isaac Bogoch @BogochIsaac · 21h

Given that Omicron is increasingly identified in countries all over the world (including in people with no connection to southern Africa), now is a good time to reconsider travel restrictions to African countries & move toward a more productive/coordinated global response.

...

All-or-nothing arguments continue to be depressingly common. Either the vaccinations 'prevent infection' or they don't, either something is safe or it isn't, and so on.

As I noted last time, travel restrictions have zero chance of *stopping* Omicron, but they will be effective at *slowing down* Omicron so long as there is a large asymmetry in Omicron's presence in different areas. I do think the time will come *relatively soon* when the restrictions stop accomplishing anything, but that time is not today.

[This](#) makes both ends of that even clearer, that it is both too late, and vital to slowing down spread if that is something you value:



Oliver Barnes @mroliverbarnes · Nov 27

Of the 600 passengers who arrived in the Netherlands yesterday from South Africa, 61 of them tested positive for Covid (Omicron not yet confirmed)

...

Then think about the below fact and that there was no testing at Heathrow yday



Oliver Barnes @mroliverbarnes · Nov 26

❗ THIS MORNING: 3 flights from South Africa arriving at London's Heathrow Airport before flight ban comes into effect at midday❗

Johannesburg - 2 flights arriving

Cape Town - 1 flight arriving

I do think it is important to show our support for South Africa, its testing and its openness. They did a heroic job and right now it's being effectively punished. What we don't want to do is make that reward 'lift the restrictions,' we want to send them other forms of aid instead, both direct help and cash and hopefully better trade and other relations, which together leaves everyone better off.

So strongly agree [with this](#):



zeynep tufekci ✅ @zeynep · 20h

Plus, I'm all for early, aggressive action when facing exponential threats but there needs to be benchmarks and a timeline for reversing precautionary steps. Such restrictions **are** a burden, and haphazardly putting a few in place and forgetting about them is not okay.



zeynep tufekci ✅ @zeynep · 19h

Replying to [@zeynep](#)

The world owes South Africa—a lot. It's also possible that they were simply the first to detect Omicron—and it's widespread already. We absolutely need to provide resources to them—and to the region. It's the right thing to do, and our moral obligation.



zeynep tufekci ✅ @zeynep · 19h

When facing an exponential threat under uncertainty, it's a **great** idea to put in friction quickly. Buying time is great! But we have to do it properly, then use that time, and be ready to reverse or change course quickly. Otherwise, we get all the burden but not the benefits.

Meanwhile, mainstream sources say [things like this](#) with straight faces.

What to know about travel after the discovery of the omicron variant

The U.S. imposed restrictions, but experts say it's too soon to cancel a trip to southern Africa

Yeah, travel bans are one thing but it is *really really not* too soon to cancel your own trip.

Merlin recommends that travelers have a back up plan, [suggesting trip stacking](#) as a potential strategy. He said travelers should follow the news closely for updates on the omicron variant and resulting restrictions — not just for Africa but for parts of the world experiencing [coronavirus](#) surges, like Europe.

I am no travel expert but your backup plan right now should very much be 'stay home for a while' or at least something not international.

Vaccinations and Boosters

If you're worried your booster shot is interfering with other people's access to the vaccines, or with 'vaccine equity,' [please stop](#). You're not doing that, except to the extent you having less and someone else not getting more means more 'equity.' [Link to article](#) Alex links to. South Africa in particular has a surplus, the same as we do.



Alex Tabarrok @ATabarrok · 22h

Replies to [@ryangrim](#) and [@literaryeric](#)

...

“South Africa has asked Johnson & Johnson (JNJ.N) and Pfizer (PFE.N) to delay delivery of COVID-19 vaccines because it now has too much stock”

If there is a new Omicron vaccine, there will once again be scarcity, but right now the bottleneck is distribution and not supply. If you’re not taking up scarce distribution, go for it.

Lockdowns

[Be very afraid of this attitude.](#)

Americans need to be prepared to do “anything and everything” to fight the omicron Covid variant, U.S. infectious disease expert Dr. Anthony Fauci said Sunday.

Still, it’s “too early to say” whether lockdowns or new mandates will be appropriate, Fauci said on ABC’s “This Week.”

Anything and everything is an appropriate response to an existential threat. I still have some probability mass in Omicron being deadlier than Delta, but almost none that it is sufficiently deadly that ‘anything and everything’ would make any sense. That kind of rhetoric is laying the foundation for very poor trade-offs whose advocates are not doing any sort of cost-benefit analysis. It’s important to be ready to push back.

Similarly, notice rhetoric where lockdowns are punishments for non-compliant populations who have sinned against public health, rather than a tool to accomplish a goal. [Here's CNN](#), note the headline about ‘the virus is in control’ framing the situation as being about who is in charge and who is or isn’t following orders properly:

The new [Omicron variant](#) might prompt a return to stricter Covid-19 measures if not enough people get vaccinated or boosted, health experts say.

...

“It ought to redouble our efforts to use the tools that we have, which are vaccinations and boosters — and to be sure we’re getting those to the rest of the world, too,” Collins told CNN on Sunday.

“It also means we need to pay attention to those mitigation strategies that people are just really sick of, like wearing masks when you’re indoors with other people who might not be vaccinated, and keeping that social distance,” he said.

...

“I think we may, indeed, be in for a phase of many more masks, much more social distancing, and more restrictions and obligations for vaccination going forward,” Schaffner said.

That doesn’t mean there are no scenarios where a *very short term* lockdown would pass a cost-benefit test to ‘flatten the curve’ or buy time for treatment supplies, but the bar on that is high and the chance of an appropriately timed response is low. Locking down or taking other extreme measures too early is quite bad, as is maintaining lockdown for too long, and conditional on lockdown I expect both.

The scariest signs are coming here in New York, [where I am worried I am going to start missing Andrew Cuomo](#) and I do not even slightly miss Andrew Cuomo.



Laurie Garrett

@Laurie_Garrett

...

ATTN New Yorkers: Figure out where, and with whom, you want to spend Winter '21-22 and get ready for another **#COVID19** hunker-down.



Kathy Hochul @GovKathyHochul · Nov 26

We continue to see warning signs of spikes in COVID this winter, and while the new Omicron variant has yet to be detected in New York State, it's coming.

Today I signed an Executive Order to help @HealthNYGov boost hospital capacity ahead of potential spikes.

[Show this thread](#)

Kathy Hochul has no idea how exponential growth works. Clearing hospital capacity *now* and stopping elective surgery *now* is exactly the *opposite* of any reasonable procedure. Even in the worst case scenarios, Omicron won't have much impact on case numbers for several weeks. After that, however long we have, we will need all the capacity we can get. The time to get other stuff out of the way is now.

The Timeline

Once again, I'll use BNO's important Omicron-related headlines in chronological order to tell the story. [Here's a link to BNO's patreon, if you'd like to help them out.](#)



BNO Newsroom @BNODesk · Nov 26

...

South Africa's health minister says, based on a small sample of Omicron cases, the majority of hospital patients are unvaccinated: "It indicates that the vaccines are providing protection"

Once again, a reminder that the majority of South Africans are unvaccinated.



BNO Newsroom @BNODesk · Nov 26

...

Dutch health service says out of some 600 people on 2 South African planes, 15 are positive for COVID-19 and 95 are negative. Nearly 500 results still pending



BNO Newsroom @BNODesk · Nov 27

Dutch health service says 61 out of 600 people on 2 flights from South Africa tested positive for coronavirus; variant not yet known

...

That last one slightly out of order, happened after the next two updates. The positive test rate went down somewhat but held at 10%, much higher than the positive test rate in South Africa according to Our World In Data. Travelers presumably are exposed more than non travelers, but still, especially since in SA who gets tested is based largely on who has symptoms.



BNO Newsroom @BNODesk · Nov 27

NEW: Germany reports probable case of new coronavirus variant in returnee from South Africa

...



BNO Newsroom @BNODesk · Nov 27

BREAKING: England reports first 2 cases of new coronavirus variant

...

It seems clear that a substantial percentage of those who fly out of South Africa are infected with Omicron, and this was mostly true the day before they started checking.



BNO Newsroom @BNODesk · Nov 27

NEW: Everyone who arrives in the UK will be required to take a PCR test on day 2 and self-isolate until a negative result is returned

...

This is the kind of half-measure that will slow things down a little but isn't pretending to want to fully work.



BNO Newsroom @BNODesk · Nov 27

Contacts of suspected Omicron cases will be required to self-isolate for 10 days, PM Johnson says

...

I am curious about the details of this.



BNO Newsroom @BNODesk · Nov 27

NEW: Israel considering to close its border to all foreigners due to new coronavirus variant

...

Boom. As I noted last time, this is the only policy with any chance of fully working.



BNO Newsroom @BNODesk · Nov 27

BREAKING: Italy reports first case of new COVID variant in traveler from Mozambique

...

I wonder what will happen when we check travelers that aren't coming from Africa?



BNO Newsroom @BNODesk · Nov 27

...

German health ministry tells BNO News that, contrary to earlier reports, 2 suspected Omicron cases have not yet been confirmed

What's weird is that it was listed as probable above?



BNO Newsroom @BNODesk · Nov 27

...

South Africa reports 3,220 new coronavirus cases, an increase of 263% from last week, with a positivity rate of 9.2%

Situation in South Africa continues to escalate super quickly, after we all started panicking so it's out-of-sample data.



BNO Newsroom @BNODesk · Nov 27

...

Israel reports 4 new suspected Omicron cases, including 3 people with no travel history

Well, ****. Presumably they found it first because they were looking.



BNO Newsroom @BNODesk · Nov 27

...

NEW: Denmark reports 2 probable cases of new COVID variant in travelers from South Africa



BNO Newsroom @BNODesk · Nov 27

Confirmed cases of new coronavirus variant:

- South Africa: 88
- Botswana: 6
- Hong Kong: 2
- England: :2
- Czech Republic: 1
- Belgium: 1
- Israel: 1
- Italy: 1



BNO Newsroom @BNODesk · Nov 28

BREAKING: Australia reports first 2 cases of new coronavirus variant

...



BNO Newsroom @BNODesk · Nov 28

BREAKING: Netherlands reports at least 13 cases of new coronavirus variant

...



BNO Newsroom @BNODesk · Nov 28

Germany reports first confirmed case of new coronavirus variant



BNO Newsroom @BNODesk · Nov 28

BREAKING: Denmark reports first 2 cases of new coronavirus variant

...



BNO Newsroom @BNODesk · Nov 28

BREAKING: Morocco bans all international flights due to new coronavirus variant

...

Morocco is an unexpected twist. Maybe worth keeping an eye on them.



BNO Newsroom @BNODesk · 23h

NEW: UK reports 3rd confirmed case of new coronavirus variant, along with "dozens" of suspected cases - FT

...



BNO Newsroom @BNODesk · 22h

NEW: Botswana reports 10 more cases of new coronavirus variant



BNO Newsroom @BNODesk · 19h

South Africa COVID update:

- New cases: 2,858
- Average: 1,975 (+310)
- Positivity rate: 9.8% (+0.6)
- In hospital: 2,232 (+3)
- In ICU: 231 (-2)
- New deaths: 6
- Average: 32 (+1)

The positivity rate continues to rise, so fall in case count presumably is about lack of testing on weekends.



BNO Newsroom @BNODesk · 17h

BREAKING: Canada reports first 2 cases of new coronavirus variant in travelers from Nigeria



BNO Newsroom @BNODesk · 17h

NEW: Austria reports 1st case of new coronavirus variant, 30 suspected cases



BNO Newsroom @BNODesk · 17h

WHO update on new coronavirus variant: "There is currently no information to suggest that symptoms associated with Omicron are different from those from other variants"



BNO Newsroom ✅ @BNODesk · 16h

WHO update on Omicron:

- Not yet known if new variant is more transmissible
- Severity of disease not yet known, no information to suggest symptoms are different
- Early evidence suggests increased risk of reinfection
- Vaccines remain critical to reduce severe disease and death

...

This is covered earlier. A lot of ‘no evidence’ talk, since it’s WHO, but at least not actively getting in the way.



BNO Newsroom ✅ @BNODesk · 11h

BREAKING: Japan closes border to all foreigners due to new coronavirus variant

...

The border closing order has a lot more to do with geography, trade and culture than suspending some flights. The contrast is interesting.



BNO Newsroom ✅ @BNODesk · 1h

Scotland reports 6 cases of new coronavirus variant, including people with no travel history, indicating community transmission

...



BNO Newsroom ✅ @BNODesk · 1h

UK investigating hundreds of suspected Omicron cases, including people who tested positive before the first known cases in Africa - Guardian

...

That last one is interesting. Going to hold off updating much on it yet, but if true then why didn’t background sequencing pick anything up?

Other New Information

There is some worry that Rapid Antigen Tests might not work on Omicron based on the sequence, but [it looks like they will continue to work](#). Presumably ‘run actual tests and see what happens’ is the way to find out, looks like we found out.

Early sign of [political reactions that were inevitable](#), but whose magnitude is not.



Kari Lake for AZ Governor @KariLake · 18h

They are going to try and sell us new “Variants” for the rest of our lives if we don’t tell them to shove it.

Some amount of this is inevitable and also necessary. There are those who *really would* have us in permanent midnight, and there’s a chance they will win, and they would happily use variants to make that happen regardless of whether it is appropriate.

Current Model

Chance that Omicron has a 100% or bigger transmission advantage in practice versus Delta: 30% → 35%.

Omicron is spreading at an alarming rate, but my confusion about what claims were being made is gone, and this is a huge transmission advantage. The counterargument is that we are seeing a lot of Omicron in a lot of places, in large numbers, very quickly, and that an advantage in Africa need not translate to a similar advantage elsewhere and we're asking here about the advantage that will hold in the United States, but also there's probably immune erosion which will matter more here than it does there. Given that this is 'in practice' which incorporates immune erosion, I'd say I'm slightly more concerned on net.

Chance that Omicron will displace Delta: 70% → 80%.

The spread and cases we are seeing clearly indicates that Omicron is spreading fast, but there have been previous variants that have taken over in some places and failed to take over other places, so that's still on the table as well. The pattern of cases we are seeing implies that Omicron is already in a lot of places, but it will take a while to be confident. Still, given my estimate of how likely it is to be immune evasive, it seems like I should be moving up my chance it eventually displaces Delta if no other variant emerges first.

Note that this probability is conditional on no other variant emerging first, and that a bunch of this probability involves a slow rather than fast transition, depending on the size of the transmission advantage.

Chance that Omicron is importantly more virulent than Delta: 25% → 10%.

Chance that Omicron is importantly less virulent than Delta: ?% → 40%

We're seeing early reports that Omicron might be *less* virulent than Delta rather than more. It didn't come directly from Delta, and also we're seeing cases show up that came from various different countries without it already having been detected, which is also evidence for it being potentially less virulent. Even more than that, it seems like evidence *against* it being *more* virulent.

Also [Zeynep points out](#) explicitly a point that's easy to miss, which is that being immune erosive will lower average severity because breakthrough cases are less severe, so Omicron will probably *look* less virulent even if it isn't.

I still think the baseline scenario is that this is similar to Delta until we hear more, but the signs are promising.

Chance that Omicron is importantly immune erosive, reducing effectiveness of vaccines and natural immunity: 50% → 80%.

This is looking likely, both in terms of what we're learning and as I get a better understanding of what our information means. The monoclonal antibodies look like they will no longer work, new versions of the vaccines are likely coming. It's still early and I don't want to get too overconfident, but yeah, this is probably happening.

Chance that Omicron means the vaccinated and previously infected are no longer effectively protected against severe disease until they get an Omicron-targeted booster shot: 5% → 4%.

Immune erosion against infection makes me more worried, but as far as I can tell that is fully consistent with protection against severe disease holding up fine. Every biological analysis and expert sees no reason to doubt that protection against severe disease will hold up.

The focus on more vaccinations now is additional evidence in favor of them holding up against severe disease, to the extent that it would vary based on the physical situation. If there were more serious doubts, I would have expected to start seeing much more loud calls for considering rushing the new versions of the vaccines and of considering holding off on additional vaccinations. That could be me *still* being naïve, though, so I'm holding onto some model uncertainty.

Chance we will be getting boosters modified for Omicron within 6 months of our previous booster shot: 15% → 30%.

They look like they are going to make Omicron booster shots, so the question is whether those already boosted will still be told to get them. Note that I consider this 'the authorities tell you to get a second booster this quickly' rather than 'I personally choose to get one this quickly.'

Chance that Omicron is less vulnerable to non-antibody treatments like Paxlovid or Fluvoxamine: 5% → 3%.

No one seems to be worried about this or have any biological basis for it, and no news is good news, so I'm going down further, but I have too much model uncertainty to go lower faster than this.

Chance we are broadly looking at a future crisis situation with widely overwhelmed American hospitals, new large American lockdowns and things like that: 20% → 15%.

There's been explicit threats and even some *current* emergency measures, which raises the chance that there will be large lockdowns whether or not they make any sense. In the other direction, the chance that this will be a deadlier variant has gone down and the chance it is less deadly has gone up. If we do get that kind of scenario, it's going to be much harder to justify extreme measures, and they would face more resistance. Also, the *really really* bad scenarios are generally less likely due to no news being good news, and I've had time to think, so I'm a little lower here.

Final Thoughts

The last few days have made it more likely that Omicron is real and is coming, although it is still too early to know for sure. They have made it more likely that immune erosion is largely responsible for Omicron's spread, as opposed to spreading more among those who haven't been infected or vaccinated. That's the bad news.

The good news is that it's more likely Omicron is milder, and less likely that Omicron is more severe. And we've reinforced that most of our treatments still work, and that vaccine protection against severe disease should hold. We're not going back to square one.

As an individual, the situation has not changed much. Be prepared for what might happen, and get ready to be more prepared as we learn more. If possible, get boosted. If you have things to do, do them *now*, while you still can, rather than prematurely locking down and doing less too early. That would only backfire.

What are the most important things to do now more broadly?

More vaccinations and booster shots help, but I believe what matters most will be our ability to:

1. Manufacture as much Paxlovid as possible as quickly as possible.
2. Get testing that will allow those with symptoms to get tested and found reliably and quickly. We still haven't fixed this. [FIX IT!](#)

3. Get a distribution system ready that's either better than or similar to the ones that exist in some states, to ensure that those who need Paxlovid and other treatments can get them quickly, without taking up too many system resources, while they will still work. If this is the kind of emergency it seems to be, once supply is adequate, you should be able to walk into a pharmacy with a Covid test you bought there, and walk out with Paxlovid, or do something *damn near* that. Again, FIX IT.

I will continue to monitor developments, and bring additional updates when it seems appropriate to do so.

Once again, I emphasize that I'm doing the best I can under uncertainty in a rapidly developing situation. I *will* make mistakes, including mistakes of thinking, and I definitely will miss important information.

I'm not looking to make any actual wagers, but I'd encourage those who have thought about things enough to have probability estimates to leave comments where you give your own current probability estimates, and explain ways in which your reasoning on them differs from mine.

The bonds of family and community: Poverty and cruelty among Russian peasants in the late 19th century

This is a linkpost for <https://rootsofprogress.org/russian-peasant-life>

Village Life in Late Tsarist Russia is an ethnographic account of Russian peasants around 1900. The author, Olga Semyonova Tian-Shanskaia ("Semyonova" for short), spent four years researching in the villages—one of the first to study a people through prolonged direct observation and contact with them.



Olga Semyonova Tian-Shanskaia

I was interested in the subject as part of learning, concretely, about the quality of life for people in various places and times and at various stages of development. Although material

progress was advancing rapidly at the end of the 19th century, much of that progress had not yet reached the province of Riazan where Semyonova did most of her studies. What was life like there?

In brief, I went in expecting poverty, which I found. I did not expect to also find a disturbing degree of cruelty and abuse.

First, let me share some passages from the book that I highlighted, roughly organized by theme. At the end I'll discuss how to interpret this and what to make of it.

In the block quotes that follow, brackets indicate insertions by the editor of the volume, David Ransel, not by me. Photographs are all taken from the Kindle edition (and presumed to be out of copyright).

Poverty



Brick masonry home, presumably in the village of Muraevnia

Here is the typical peasant diet:

Dinner consists of cabbage soup, porridge (*kasha*), or, again, potatoes and bread. The soup, of course, has no meat in it, just cabbage, but occasionally sour cream is added. Potatoes are mixed with kvass and onions. The porridge—usually millet—is eaten either with milk, in what is known as the thin form (*kulesh*), or in a thicker form made with hemp-seed oil. After dinner, the peasants take a rest and then return to the fields. They take along some bread for an afternoon snack sometime between three and five o'clock. When the sun sets, the peasants go home, and at about nine in the evening they eat their supper, which is warmed-over dinner, with the possible addition of skim milk.

You get a better sense of the standard of living from the foods that they consider special treats:

Foods that can be found on a peasant's table only during the annual festival, or during a wedding or baptismal dinner, include pancakes (*bliny*), meat (veal designated for holidays—*uboina*), potato fritters (*drachena*), buns (*pyshki*), *salamata* (a kind of thin gruel), *kalinnik* (a kind of cake), fritters (*olad'i*), and cabbage soup with corned beef (*solonina*).

That, of course, is during good times. In bad times:

During a famine, peasant meals consist of stale bread moistened in water and mixed with goosefoot. The men make extra efforts to find any type of work, and sometimes the entire family goes out to beg. As soon as the snow melts, hungry children pick roots and herbs to eat, such as sorrel and clover. Peasants also make soup with goutweed (*Aegopodium podagraria*).

Another hallmark of poverty is the burning of solid fuels inside the home. Here's a visualization of the problems that causes:

Stoves with chimneys are called "white," while chimneyless stoves are referred to as "black." When a "black" stove is being lighted, the door from the main room into the entryway is left open so that the smoke up to the level of the door is drawn out, but above that level it forms a blue and white blanket through which nothing can be seen. The top of the door is no higher than a rather short person, so that one has to stoop to enter the house (the ceiling itself is only 5'10" high). A good-sized man finds it difficult to stand up when the stove is being fired, because his eyes will be in the caustic cloud of smoke. Even when seated on a bench, one feels the acrid smoke in the eyes.

It's bad enough to burn wood and coal; these peasants often had to burn straw, or worse:

Peasants bring straw in at night to sleep on, and then in the morning they feed this straw into the stove. The use of fresh straw each day provides reasonably hygienic bedding for the peasants. But this is so only in years of abundant harvest. In a bad year, whatever little straw they have is given to the livestock; sometimes even straw roofs have to be pulled down to save the animals from starvation.

The shortage of straw forces peasants to use their clothes for bedding and to heat the house with dried manure or weeds such as burdock, thistle, and nettles. Accordingly, illnesses increase in such a year. The lack of fresh bedding is one cause. The poor fuel likewise does much damage to the eyes. In the drought years of 1891–1892, around ten people in two of our small villages (each containing about fifteen households) lost their eyesight temporarily or permanently from the smoke of their stoves. The smoke, which was produced by burning dried manure and weeds found on the roadside and in ravines, was so acrid that the victims (mostly old people and children) developed cataracts. All of them were admitted to the regional hospital in town, but three of them never got their eyesight back.

Unfortunately, the peasants were also ignorant of basic hygiene—in the Tsarist era, the germ theory had not yet made its way from the Institut Pasteur to the Russian peasantry:

If a mouse falls into a tub of pickles, sauerkraut, or pickled apples, the woman of the house will nearly always summon the priest to perform a cleansing ritual. The mouse is plucked from the tub, and the priest proceeds to say a prayer over the tub, pass a cross over it three times, and then bite into a pickle or apple or try some of the sauerkraut. After this, the contents of the tub are again regarded as clean.

And the combination of this ignorance with poverty resulted in health disasters:

In the work season, peasants often catch colds and lose their voices ("My whole chest is blocked up," they report) because they drink cold water when they are overheated.Flushed with heat and "assaulted by thirst," as the peasants say, they cannot seem to get enough to drink and will take water from any source available. They drink from a roadside ditch, from a muddy puddle, from a swamp, wherever they can find some water to quench their thirst. Tapeworms and roundworms are common among peasants. In the fall, the water in the river is literally poisoned by hemp that is soaked there until the first frosts. There are cases of poisoning with this water.

Morals



Peasant man making barrel hoops, Riazan province. Riazan Museum

Semyonova is very unimpressed with Russian peasants' work ethic:

... peasants give no thought to saving for the future. If a peasant has a good harvest that will keep him until the next year, he will stay at home and loaf and cannot be enticed to take an extra job at any price. Peasants hire themselves out as farm laborers only when driven to it by dire need, when they have, so to speak, a knife at their throat. [Then they will do] the most arduous work imaginable.

The peasants' lack of respect for hard work is remarkable. "Him? He digs in the field like a beetle from morning till night!" They often say this with scorn.

She also saw resentment and envy:

The better-off peasants are bitter about the attitude of their poorer neighbors. "They hate and envy us constantly, saying things like: 'What makes you think you're so much better? Just wait, you're going to be as poor as us.' If you plant an apple tree, they resent it, saying: 'Now that big shot is planting an orchard! We are starving while he is putting in an orchard, and fencing it off at that!'" And they think nothing of breaking down the fence and uprooting the tree. If the tree happens to survive and bear fruit, they feel it is their duty to raid it. "That's how much hate they have! And if some misfortune should befall you, they'll make sure to finish you off."

True to the Russian stereotype, the peasants drank a lot:

Prodigious amounts of alcohol are consumed at wedding parties. I myself have attended weddings at which nine-and ten-year-old girls were made to drink so that they would dance for everyone's entertainment.

The best occasion for young people to get drunk for the first time is the annual festival, which in this region takes place in connection with St. Michael's Day. On that holiday, every person in the parish is drunk. In a good year the festival lasts for a week, but even when the crops are poor, people manage to go on a spree for three days.

Accidents abound in the springtime [as a result of drunkenness]. Some people drown in water-filled ravines. Others are crushed under falling wagons; a drunken peasant will have a wagon tip over on him, and that is the end of him.

Another occasion for general drunkenness is seasonal field work for the landlord (usually mowing and transportation of produce to town), who by way of payment treats the peasants to refreshments. On these occasions dreadful fights break out and can result in maiming or even killing with a scythe.

After the [church] service, abstemious peasants go home, while others head for the tavern, where everyone gets drunk. The more sober peasants eat dinner at home, rest, and then "just sit" till the evening, talking about their affairs and discussing the harvest and related subjects. Yet even back in the home village, there are opportunities to have a drink, and it is unlikely that a peasant will let a Sunday go by without having one. In the evening, women get a feeling for the amount of alcohol consumed by their husbands by the intensity of the beatings they receive.

Also true to the stereotype:

Fistfighting and swearing are learned quite early. As soon as Ivan began to walk, he started fighting with other children. He was actually encouraged to do this, especially if he was able to best another small child. Ivan learned swear words from his older brothers and sisters, even before he could put together a complete sentence. He started to call his mother a bitch whenever she denied him something, much to the delight of the whole family, even the mother herself. They would actually encourage him on such occasions.

("Ivan" here is a composite figure representing the typical Russian peasant child.)

Theft was common:

When a group of young horseherds includes a few older boys, say about age sixteen or seventeen, they instruct the younger ones to steal liquor from home when their parents are away. The loot is then shared by all, not infrequently including boys ten to twelve years old.

Theft between the spouses is not uncommon. It may be the husband who steals money from his wife's trunk for some "need" of his or just to enjoy himself at a tavern. Or the wife might take some flour or grain from her husband and use it to pay for soap or some satin cloth at the store. When a husband is drunk, his wife will slip his wallet out of his boot. Children, too, steal eggs or anything handy from their mothers. Wives swipe wool from their husbands.

In one incident, Semyonova hired a peasant woman to cook, and gave her flour to make bread. The woman used some of the flour to bake cakes for herself and her husband. Semyonova talked to another of the servants, Katerina, about this:

Katerina: "That's not stealing! She just baked and ate it. She didn't take the cakes to her room or hide them in the storeroom."

I: "But she took the flour, and as a result the laborers had less bread. What's the difference if she stole it and took it home or ate it right here? It's still robbery."

Katerina: "But she ate the cakes right here in your house, together with her husband; that's not robbery. If she had stolen the flour from a locked cupboard or saved it for the future, that would probably be a sin."

Try as I might to explain to Katerina that unauthorized appropriation of another person's property, whether consumed immediately or saved for future use, is still a theft, she would not agree with me.

Semyonova notes that "the very same elder" who turned in the flour thief, "when he is guarding the landlord's apple trees against raids by the boys hired on temporarily as shepherds, fills his pockets with apples every time he makes the rounds."

Cruelty



*Peasant girls from the village of Kultuki,
Kasimov district of Riazan province. Riazan
Museum*

Warning: from here on out, this post may be difficult to read.

The peasants had no regard at all for animals:

Cats and dogs are also less useful than other animals, and peasants will torture them just for the fun of it, just to see what will happen. Little children like to throw cats and puppies, when they can catch them, into the water to see if they can swim. When I ask, "Don't you feel sorry for them?" the children respond: "Why feel sorry? They're not people, just dogs."

I cannot say that peasants treat their livestock especially well. Horses are routinely beaten. Yet a peasant feels very upset if a horse dies, because this is a great financial loss. A woman reacts the same way to the loss of a cow.

Worse, they also abused their children:

Punishment for mischief consisted of beatings administered by the parents. They beat Ivan for screaming, getting covered with mud, or stealing a piece of food. They did not beat him for fighting, lying, or using foul language.

We also continue to see cases of sons being flogged for insulting their parents. These disobedient children (who may be as much as twenty years of age) are taken without trial to the township office for their punishment. In response to a complaint by the parents, the township supervisor summons the son and turns him over to the office guard for flogging. The culprit is stripped from the waist down, placed on the floor in the township office, and beaten with willow rods. Admission to the spectacle is open to all residents of the village.

There was even mention of a case of child rape:

There was a case this summer in which a twenty-year-old guard at the apple orchard raped a thirteen-year-old girl. The mother of the girl—a very poor woman, it is true—agreed to forgive the offender in exchange for three rubles.

I'm not sure which is more shocking: that the offender got off with a fine, or that whether to forgive him was the mother's decision.

Women



A group of peasant women. Library of Congress

More generally, women at all stages of life seemed to be treated very badly—starting from birth:

But when a girl was born, the grandparents stopped thinking about her as soon as she was baptized. They did not even express any sorrow about her death. The young father, too, did not feel much regret over it.

If the first child is a girl, the feeling in the family is mostly one of disappointment. One of the women might remark: "Oh well, at least she can be a nursemaid." By the following day, no one gives a thought to the baby girl.

Here, for instance, is what happens at the "bride-show," which happens one to two weeks after a marriage proposal:

The bride is now standing in front of the groom. Her head kerchief is tied in such a way that her face is shaded. The groom's family inspects the bride closely. "Perhaps she is lame?" they may inquire. Her sister or sister-in-law then has her walk around the room. Other questions are asked, for example, about whether she may be deaf. Then her future parents-in-law approach the bride and ask: "Why is she wrapped like that so we can't see

her eyes; is she possibly blind in one eye?" The bride's sister uncovers her face for inspection. If the bride is pale, the groom's family will want to know the reason, asking whether she is sickly.

If you are thinking that this does not augur well for the marriage, the bride and her family would agree with you: later, back home, the bride "pounds her head on a bench while crying out a lamentation or, as it is known here, 'the scream.' Her mother and sister join in:"

My father, my provider,
My dearest mother,
I've been given away, miserable and hapless,
And giving me away they washed it down with vodka
And a burnt bread crust.
How will it turn out for me going to live with strangers,
To a new father and mother.
I will have to please these strangers,
To be pleasing and obedient to them all.

Indeed, the marriage does not get off to a good start. After the wedding:

In the morning, the best man and the godmother wake up the bridal couple. The godmother orders the young wife to sweep the floors. Copper coins have been tossed around on the floor beforehand, and the wife is told to give her mother-in-law any coins she finds. This is done to find out if the young wife is a thief, and also to see how well she sweeps the floor.

Nor are women supported well in pregnancy:

During pregnancy, a woman continues to be responsible for all her usual chores, both in the household and in the field—including binding the sheaves, weeding, threshing, gathering in the hemp, planting and digging potatoes—right up to the onset of labor. Women frequently give birth while performing a domestic chore, such as kneading bread, or even when they are at work in the field; others do so riding in a bumpy wagon as they are hurrying home after being prompted by the first pangs of the approaching birth.

Infants



*Young mother holds her swaddled newborn in
the Sapozhok district of the Riazan province.
Riazan Museum*

But most shocking to me—again, this will be difficult to read—was the treatment of infants:

Mothers who are concerned about neatness put straw into the cradle and change it every day or two. More often, though, the baby is placed into a dirty cradle lined with its mother's soiled old skirt: "He can just as well lie on the skirt, no better than anybody else. Others didn't seem to die; they survived."

Up to the time Ivan takes his first steps, he is looked after by his sister, a girl of nine or ten years of age. She has difficulty carrying him around and often drops him, exclaiming: "Oops, my goodness! How did I let go of him?" Sometimes Ivan tumbles headfirst down a hillock. When he cries, his baby-sitter uses her free hand to slap him on the face or head, saying, "Keep quiet, you son of a bitch."

Mothers sometimes sang cruel lullabies, such as:

Hush, hush, hushaby my baby,
I'll give you spankings,
Twenty-five of them,
To make you sleep better and deeper.

Hush, hush, hushaby my baby.
A man lives at the end of the village.
He's neither poor, nor rich,
He has many children,
They sit on a bench
And eat straw.
I'll make you suffer even more.
I won't give you anything to eat.
I won't make a bed for you.

And some were even worse than these; a footnote added by the editor says:

A Russian folklorist found that about 8 percent of her collection of thousands of lullabies were songs wishing death on babies, presumably weak infants like those mentioned here whose survival was uncertain and who may have been in pain.

Indeed, it seems that the death of a child was not always considered a tragedy:

[The death of an infant in a poor family that could not support another child was evidently regarded as a blessing, as Semyonova recorded in one of her unpublished field notes.] When a poor family's child dies, people say: "Thank goodness, the Lord thought better of it!"

And child mortality was high. Some deaths were the result of poverty and lack of hygiene:

The rate of child deaths is highest in the summer during the fast of St. Peter [in June], and especially during the field-work season, when unattended children eat anything they come across: cucumbers, sour apples, and any other vegetation. Diarrhea is the chief cause of child death. As for the death rate, in a majority of homes more than half of all children die. Most women bear from eight to ten or twelve children, of which only three or four survive.

Some were due to smothering in bed, which was claimed to be accidental, but which happened so often that many observers believed it to be deliberate infanticide:

Moreover, young mothers very often smother their children accidentally in their sleep. The mother sometimes places her infant between herself and her husband to give the baby her breast, goes to sleep, rolls over on the baby, and smothers it. A good half of the women have overlain at least one child in this way—they do it most often in their young years when they sleep soundly. For overlying a child, the priest imposes a penance.

(A penance!)

And in the case of illegitimate children, even the pretense of accident was dropped:

Cases of infanticide of illegitimate babies are not at all rare. A married or unmarried woman gives birth alone somewhere in a shed, smothers the baby, and dumps it into the river (with a rock secured to its neck) or leaves it in a hemp thicket, or buries it either in the yard or somewhere in the pigpen.

The parents of a young woman who had gotten pregnant out of wedlock married her off to hide her sin. When the woman gave birth, her husband's family [with whom she was then living] turned against the child. Although her husband was a peaceful, simple-hearted fellow who did not reproach his wife for her youthful indiscretion, his family was relentless and eventually demanded that she "get rid of the little bastard." This demand was so insistent, the poor woman being continually beaten and persecuted by her in-laws, that she gave in. She filled the infant's pacifier rag with sulfur scraped off matches, placed it in the baby's mouth, and it soon died. The mother was taken to court but was acquitted.

Not all babies are subject to killing, but illegitimate ones are very likely to be. A weak baby that is a burden to its mother is not killed, although its parents will grumble at its existence and constantly express a desire for its death.

Older women are very ruthless and cold-blooded about the killing of an illegitimate "whelp," whom they view as a nuisance and a burden. Young women anguish over such a decision and force themselves to kill their babies only when shame or fear causes them to lose their senses, or when they can no longer bear their own and their babies' suffering. Men, from what I could observe, often simply do not know about such killings, even when they occur in their own families, or, if they figure out what is going on, they look the other way.

The editor adds:

Semyonova's observations are disturbing to modern readers, but there is no reason to doubt their accuracy. My own recent researches on this question suggest that in many parts of Russia, children, unwanted because of illegitimacy, physical malformation, or apparent weakness, met their deaths either quickly through infanticide of the kinds described by Semyonova or more slowly by a reduced level of care and feeding.

What to make of all this?

First, note that this book is not entirely the work of Semyonova or even of her collaborator, K. V. Nikolaevskii. She left the work unfinished when she died, and her notes were put together into a coherent form by Ransel, the editor.

Second, Semyonova was not an entirely neutral observer. As Ransel says in the introduction:

She clearly sensed the otherness of the peasant world and regarded the peasants as different from educated, urbanized Russians in fundamental ways. ... Her stance was that of a progressive, westernizing member of the Russian intelligentsia. She wanted the peasants to share her respect for private property, her values of thrift and hard work, and she believed that without these values they could not become enlightened and productive citizens of a modern society. ... To sum up, even though Semyonova renders marvelously detailed and striking portrayals of peasant behavior, she makes little effort to enter the peasants' cultural frame. She remains outside and is entirely confident of her superiority to the peasants.

Further, she had an agenda, or at least a specific cultural purpose. Ransel quotes a 1906 letter from her collaborator Nikolaevskii: "The political and social climate of that time (as is still true today) was such that everyone expected that the peasant alone would be able to bring about a new order in Russia, and the pace of this change was also entirely dependent on the peasantry." Ransel says, "Clearly, a central purpose of her study is to counter naive views of the peasants as naturally cooperative, communitarian beings who will provide the foundation for a new order of social peace and harmony."

All that said, my estimation is that the specific incidents and at least the first-level generalizations are true, even if they have to be interpreted from Semyonova's "progressive, westernizing" frame.

The question for me is how much these observations apply to peasant life in other places and at other times. I'm hesitant to generalize, since this is the first book-length work of ethnography I've read in the context of this project, but for me it opens questions. Is cruelty towards animals and children, and an almost slave status for women, the norm? If Russian peasants were more cruel than average, are they *far* worse than average? If this kind of cruelty is common, is it inherent to poverty, including the lower levels of education that necessarily accompany poverty? And if so, does increased wealth and education alone lead to a more humane society, or did the transition to more equal rights and status also require a change in morals and other ideas that is not a natural or inherent consequence of material progress?

The main reason this matters is that equal rights and humane treatment are part of human well-being, and therefore how those things were achieved is part of the story of progress. But a secondary reason I am interested is that one criticism of modernity claims that today we are more "disconnected" from our families and our communities. In the case of Russian peasant women in particular, their "connection" to husbands who beat them, in-laws who scorned and humiliated them, and a community that offered no support, was not enriching but immiserating. The escape from those "connections," provided by wealth, education, and opportunities for jobs and migration, would have been a boon.

Ngo and Yudkowsky on AI capability gains

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

This is the second post in a series of transcribed conversations about AGI forecasting and alignment. See the [first post](#) for prefaces and more information about the format.

Color key:

Chat by Richard Ngo and Eliezer Yudkowsky

Other chat

Inline comments

5. September 14 conversation

5.1. Recursive self-improvement, abstractions, and miracles

[Yudkowsky][11:00]

Good morning / good evening.

So it seems like the obvious thread to pull today is your sense that I'm wrong about recursive self-improvement and consequentialism in a related way?

[Ngo][11:04]

Right. And then another potential thread (probably of secondary importance) is the question of what you mean by utility functions, and digging more into the intuitions surrounding those.

But let me start by fleshing out this RSI/consequentialism claim.

I claim that your early writings about RSI focused too much on a very powerful abstraction, of recursively applied optimisation; and too little on the ways in which even powerful abstractions like this one become a bit... let's say messier, when they interact with the real world.

In particular, I think that [Paul's arguments](#) that there will be substantial progress in AI in the leadup to a RSI-driven takeoff are pretty strong ones.

(Just so we're on the same page: to what extent did those arguments end up shifting your credences?)

[Yudkowsky][11:09]

I don't remember being shifted by Paul on this at all. I sure shifted a lot over events like Alpha Zero and the entire deep learning revolution. What does Paul say that isn't encapsulated in that update - does he furthermore claim that we're going to get fully smarter-than-human in all regards AI which doesn't cognitively scale much further either through more compute or through RSI?

[Ngo][11:10]

Ah, I see. In that case, let's just focus on the update from the deep learning revolution.

[Yudkowsky][11:12][11:13]

I'll also remark that I see my foreseeable mistake there as having little to do with "abstractions becoming messier when they interact with the real world" - this truism tells you very little of itself, unless you can predict *directional* shifts in other variables just by contemplating the *unknown* messiness relative to the abstraction.

Rather, I'd see it as a neighboring error to what I've called the Law of Earlier Failure, where the Law of Earlier Failure says that, compared to the interesting part of the problem where it's fun to imagine yourself failing, you usually fail before then, because of the many earlier boring points where it's possible to fail.

The nearby reasoning error in my case is that I focused on an interesting way that AI capabilities could scale and the most powerful argument I had to overcome Robin's objections, while missing the way that Robin's objections could fail even earlier through rapid scaling and generalization in a more boring way.

It doesn't mean that my arguments about RSI were false about their domain of supposed application, but that other things were also true and those things happened first on our timeline. To be clear, I think this is an important and generalizable issue with the impossible task of trying to forecast the Future, and if I am wrong about other things it sure would be plausible if I was wrong in similar ways.

[Ngo][11:13]

Then the analogy here is something like: there is a powerful abstraction, namely consequentialism; and we both agree that (like RSI) a large amount of consequentialism is a very dangerous thing. But we disagree on the question of how much the strategic landscape in the leadup to highly-consequentialist AIs is affected by other factors apart from this particular abstraction.

"this truism tells you very little of itself, unless you can predict directional shifts in other variables just by contemplating the unknown messiness relative to the abstraction"

I disagree with this claim. It seems to me that the predictable direction in which the messiness pushes is *away from* the applicability of the high-level abstraction.

[Yudkowsky][11:15]

The real world is messy, but good abstractions still apply, just with some messiness around them. The Law of Earlier Failure is not a failure of the abstraction being messy, it's a failure of the *subject matter* ending up different such that the abstractions you used were *about a different subject matter*.

When a company fails before the exciting challenge where you try to scale your app across a million users, because you couldn't hire enough programmers to build your app at all, the problem is not that you had an unexpectedly messy abstraction about scaling to many users, but that the key determinants were a different subject matter than "scaling to many users".

Throwing 10,000 TPUs at something and actually getting progress - not very much of a famous technological idiom *at the time I was originally arguing with Robin* - is not a leak in the RSI abstraction, it's just a way of getting powerful capabilities without RSI.

[Ngo][11:18]

To me the difference between these two things seems mainly semantic; does it seem otherwise to you?

[Yudkowsky][11:18]

If I'd been arguing with somebody who kept arguing in favor of faster timescales, maybe I'd have focused on that different subject matter and gotten a chance to be explicitly wrong about it. I mainly see my ur-failure here as letting myself be influenced by the whole audience that was nodding along very seriously to Robin's arguments, at the expense of considering how reality might depart in either direction from my own beliefs, and not just how Robin might be right or how to persuade the audience.

[Ngo][11:19]

Also, "throwing 10,000 TPUs at something and actually getting progress" doesn't seem like an example of the Law of Earlier Failure - if anything it seems like an Earlier Success

[Yudkowsky][11:19]

it's an Earlier Failure of Robin's arguments about why AI wouldn't scale quickly, so my lack of awareness of this case of the Law of Earlier Failure is why I didn't consider why Robin's arguments could fail earlier

though, again, this is a bit harder to call if you're trying to call it in 2008 instead of 2018

but it's a valid lesson that the future is, in fact, hard to predict, if you're trying to do it in the past

and I would not consider it a merely "semantic" difference as to whether you made a wrong argument about the correct subject matter, or a correct argument about the wrong subject matter

these are like... very different failure modes that you learn different lessons from

but if you're not excited by these particular fine differences in failure modes or lessons to learn from them, we should perhaps not dwell upon that part of the meta-level Art

[Ngo][11:21]

Okay, so let me see if I understand your position here.

Due to the deep learning revolution, it turned out that there were ways to get powerful capabilities without RSI. This isn't intrinsically a (strong) strike against the RSI abstraction; and so, unless we have reason to expect another similarly surprising revolution before reaching AGI, it's not a good reason to doubt the consequentialism abstraction.

[Yudkowsky][11:25]

Consequentialism and RSI are very different notions in the first place. Consequentialism is, in my own books, significantly simpler. I don't see much of a conceptual connection between the two myself, except insofar as they both happen to be part of the connected fabric of a coherent worldview about cognition.

It is entirely reasonable to suspect that we may get another surprising revolution before reaching AGI. Expecting a *particular* revolution that gives you *particular* miraculous benefits is much more questionable and is an instance of conjuring expected good from nowhere, like hoping that

you win the lottery because the first lottery ball comes up 37. (Also, if you sincerely believed you actually had info about what kind of revolution might lead to AGI, you should shut up about it and tell very few carefully selected people, not bake it into a public dialogue.)

[Ngo][11:28]

and I would not consider it a merely "semantic" difference as to whether you made a wrong argument about the correct subject matter, or a correct argument about the wrong subject matter

On this point: the implicit premise of "and also nothing else will break this abstraction or render it much less relevant" turns a correct argument about the wrong subject matter into an incorrect argument.

[Yudkowsky][11:28]

Sure.

Though I'd also note that there's an important lesson of technique where you learn to say things like that out loud instead of keeping them "implicit".

Learned lessons like that are one reason why I go through your summary documents of our conversation and ask for many careful differences of wording about words like "will happen" and so on.

[Ngo][11:30]

Makes sense.

So I claim that:

1. A premise like this is necessary for us to believe that your claims about consequentialism lead to extinction.
2. A surprising revolution would make it harder to believe this premise, even if we don't know which *particular* revolution it is.
3. If we'd been told back in 2008 that a surprising revolution would occur in AI, then we should have been less confident in the importance of the RSI abstraction to understanding AGI and AGI risk.

[Yudkowsky][11:32][11:34]

Suppose I put to you that this claim is merely subsumed by all of my previous careful qualifiers about how we might get a "miracle" and how we should be trying to prepare for an unknown miracle in any number of places. Why suspect that place particularly for a model-violation?

I also think that you are misinterpreting my old arguments about RSI, in a pattern that matches some other cases of your summarizing my beliefs as "X is the one big ultra-central thing" rather than "X is the point where the other person got stuck and Eliezer had to spend a lot of time arguing".

I was always claiming that RSI was a way for AGI capabilities to scale much further *once they got far enough*, not the way AI would scale to *human-level generality*.

This continues to be a key fact of relevance to my future model, in the form of the unfalsified original argument about the subject matter it previously applied to: if you lose control of a sufficiently smart AGI, it will Foom, and this fact about what triggers the metaphorical equivalent of a full nuclear exchange and a total loss of the gameboard continues to be extremely relevant to what you have to do to obtain victory instead.

[Ngo][11:34][11:35]

Perhaps we're interpreting the word "miracle" in quite different ways.

I think of it as an event with negligibly small probability.

[Yudkowsky][11:35]

Events that actually have negligibly small probability are not much use in plans.

[Ngo][11:35]

Which I guess doesn't fit with your claims that we should be trying to prepare for a miracle.

[Yudkowsky][11:35]

Correct.

[Ngo][11:35]

But I'm not recalling off the top of my head where you've claimed that.

I'll do a quick search of the transcript

"You need to hold your mind open for any miracle and a miracle you didn't expect or think of in advance, because at this point our last hope is that in fact the future is often quite surprising."

Okay, I see. The connotations of "miracle" seemed sufficiently strong to me that I didn't interpret "you need to hold your mind open" as practical advice.

What sort of probability, overall, do you assign to us being saved by what you call a miracle?

[Yudkowsky][11:40]

It's not a place where I find quantitative probabilities to be especially helpful.

And if I had one, I suspect I would not publish it.

[Ngo][11:41]

Can you leak a bit of information? Say, more or less than 10%?

[Yudkowsky][11:41]

Less.

Though a lot of that is dominated, not by the probability of a positive miracle, but by the extent to which we seem unprepared to take advantage of it, and so would not be saved by one.

[Ngo][11:41]

Yeah, I see.

5.2. The idea of expected utility

[Ngo][11:43]

Okay, I'm now significantly less confident about how much we actually disagree.

At least about the issues of AI cognition.

[Yudkowsky][11:44]

You seem to suspect we'll get a *particular* miracle having to do with "consequentialism", which means that although it might be a miracle to me, it wouldn't be a miracle to you.

There is something forbidden in my model that is not forbidden in yours.

[Ngo][11:45]

I think that's partially correct, but I'd call it more a *broad range of possibilities* in the rough direction of you being wrong about consequentialism.

[Yudkowsky][11:46]

Well, as much as it may be nicer to debate when the other person has a specific positive expectation that X will work, we can also debate when I know that X won't work and the other person remains ignorant of that. So say more!

[Ngo][11:47]

That's why I've mostly been trying to clarify your models rather than trying to make specific claims of my own.

Which I think I'd prefer to continue doing, if you're amenable, by asking you about what entities a utility function is defined over - say, in the context of a human.

[Yudkowsky][11:51][11:53]

I think that to contain the concept of Utility as it exists in me, you would have to do homework exercises I don't know how to prescribe. Maybe one set of homework exercises like that would be showing you an agent, including a human, making some set of choices that allegedly couldn't obey expected utility, and having you figure out how to pump money from that agent (or present it with money that it would pass up).

Like, just actually doing that a few dozen times.

Maybe it's not helpful for me to say this? If you say it to Eliezer, he immediately goes, "Ah, yes, I could see how I would update that way after doing the homework, so I will save myself some time and effort and just make that update now without the homework", but this kind of jumping-ahead-to-the-destination is something that seems to me to be... dramatically missing from many non-Eliezers. They insist on learning things the hard way and then act all surprised when they do. Oh my gosh, who would have thought that an AI breakthrough would suddenly make AI seem less than 100 years away the way it seemed yesterday? Oh my gosh, who would have thought that alignment would be difficult?

Utility can be seen as the origin of Probability within minds, even though Probability obeys its own, simpler coherence constraints.

that is, you will have money pumped out of you, unless you weigh in your

mind paths through time according to some quantitative weight, which determines how much resources you're willing to spend on preparing for them

this is why sapients think of things as being more or less likely

[Ngo][11:53]

Suppose that this agent has some high-level concept - say, honour - which leads it to pass up on offers of money.

[Yudkowsky][11:55]

Suppose that this agent has some high-level concept - say, honour - which leads it to pass up on offers of money.

then there's two possibilities:

- this concept of honor is something that you can see as helping to navigate a path through time to a destination
- honor isn't something that would be optimized into existence by optimization pressure for other final outcomes

[Ngo][11:55]

Right, I see.

Hmm, but it seems like humans often don't see concepts as helping to navigate a path in time to a destination. (E.g. the deontological instinct not to kill.)

And yet those concepts were in fact optimised into existence by evolution.

[Yudkowsky][11:59]

You're describing a defect of human reflectivity about their consequentialist structure, not a departure from consequentialist structure. 😊

[Ngo][12:01]

(Sorry, internet was slightly buggy; switched to a better connection now.)

[Yudkowsky][12:01]

But yes, from my perspective, it creates a very large conceptual gap that I can stare at something for a few seconds and figure out how to parse it

as navigating paths through time, while others think that "consequentialism" only happens when their minds are explicitly thinking about "well, what would have this consequence" using language.

Similarly, when it comes to Expected Utility, I see that any time something is attaching relative-planning-weights to paths through time, not when a human is thinking out loud about putting spoken numbers on outcomes

[Ngo][12:02]

Human consequentialist structure was optimised by evolution for a different environment. Insofar as we are consequentialists in a new environment, it's only because we're able to be reflective about our consequentialist structure (or because there are strong similarities between the environments).

[Yudkowsky][12:02]

False.

It just generalized out-of-distribution because the underlying coherence of the coherent behaviors was simple.

When you have a very simple pattern, it can generalize across weak similarities, not "strong similarities".

The human brain is large but the coherence in it is simple.

The idea, the structure, that explains why the big thing works, is much smaller than the big thing.

So it can generalize very widely.

[Ngo][12:04]

Taking this example of the instinct not to kill people - is this one of the "very simple patterns" that you're talking about?

[Yudkowsky][12:05]

"Reflectivity" doesn't help per se unless on some core level a pattern already generalizes, I mean, either a truth can generalize across the data or it can't? So I'm a bit puzzled about why you're bringing up "reflectivity" in this context.

And, no.

An instinct not to kill doesn't even seem to me like a plausible cross-cultural universal. 40% of deaths among Yanomami men are in intratribal

fights, iirc.

[Ngo][12:07]

Ah, I think we were talking past each other. When you said "this concept of honor is something that you can see as helping to navigate a path through time to a destination" I thought you meant "you" as in the agent in question (as you used it in some previous messages) not "you" as in a hypothetical reader.

[Yudkowsky][12:07]

ah.

it would not have occurred to me to ascribe that much competence to an agent that wasn't a superintelligence.

even I don't have time to think about why more than ~~0.0001%~~ 0.01% of my thoughts do anything, but thankfully, you don't have to think about *why* $2 + 2 = 4$ for it to be the correct answer for counting sheep.

[Ngo][12:10]

Got it.

I might now try to throw a high-level (but still inchoate) disagreement at you and see how that goes. But while I'm formulating that, I'm curious what your thoughts are on where to take the discussion.

Actually, let's spend a few minutes deciding where to go next, and then take a break

I'm thinking that, at this point, there might be more value in moving onto geopolitics

[Yudkowsky][12:19]

Some of my current thoughts are a reiteration of old despair: It feels to me like the typical Other within EA has no experience with discovering unexpected order, with operating a generalization that you can expect will cover new cases even when that isn't immediately obvious, with operating that generalization to cover those new cases correctly, with seeing simple structures that generalize a lot and having that be a real and useful and technical experience; instead of somebody blathering in a non-expectation-constraining way about how "capitalism is responsible for everything wrong with the world", and being able to extend that to lots of cases.

I could try to use much simpler language in hopes that people actually [look-at-the-water](#) Feynman-style, like "navigating a path through time"

instead of Consequentialism which is itself a step down from Expected Utility.

But you actually do lose something when you throw away the more technical concept. And then people still think that either you instantly see in the first second how something is a case of "navigating a path through time", or that this is something that people only do explicitly when visualizing paths through time using that mental terminology; or, if Eliezer says that it's "navigating time" anyways, this must be an instance of Eliezer doing that thing other people do when they talk about how "Capitalism is responsible for all the problems of the world". They have no experience operating genuinely useful, genuinely deep generalizations that extend to nonobvious things.

And in fact, being able to operate some generalizations like that is a lot of how I know what I know, in reality and in terms of the original knowledge that came before trying to argue that knowledge with people. So trying to convey the real source of the knowledge feels doomed. It's a kind of idea that our civilization has lost, like that college class Feynman ran into.

[Soares][12:19]

My own sense (having been back for about 20min) is that one of the key cruxes is in "is it possible that non-scary cognition will be able to end the acute risk period", or perhaps "should we expect a longish regime of pre-scary cognition, that we can study and learn to align in such a way that by the time we get scary cognition we can readily align it".

[Ngo][12:19]

Some potential prompts for that:

- what are some scary things which might make governments take AI more seriously than they took covid, and which might happen before AGI
- how much of a bottleneck in your model is governmental competence? and how much of a difference do you see in this between, say, the US and China?

[Soares][12:20]

I also have a bit of a sense that there's a bit more driving to do on the "perhaps EY is just wrong about the applicability of the consequentialism arguments" (in a similar domain), and would be happy to try articulating a bit of what I think are the not-quite-articulated-to-my-satisfaction arguments on that side.

[Yudkowsky][12:21]

I also had a sense - maybe mistaken - that RN did have some *specific* ideas about how "consequentialism" might be inapplicable, though maybe I accidentally refuted that in passing because the idea was "well, what if it didn't know what consequentialism was?" and then I explained that reflectivity was not required to make consequentialism generalize. but if so, I'd like RN to say explicitly what specific idea got refuted that way, or failing that, talk about the specific idea that didn't get refuted.

[Ngo][12:23]

That wasn't my objection, but I do have some more specific ideas, which I could talk about.

And I'd also be happy for Nate to try articulating some of the arguments he mentioned above.

[Yudkowsky][12:23]

I have a general worry that this conversation has gotten too general, and that it would be more productive, even of general understanding, to start from specific ideas and shoot those down specifically.

[Ngo: 

[Ngo][12:26]

The other thing is that, for pedagogical purposes, I think it'd be useful for you to express some of your beliefs about how governments will respond to AI

I think I have a rough guess about what those beliefs are, but even if I'm right, not everyone who reads this transcript will be

[Yudkowsky][12:28]

Why would I be expected to know *that*? I could talk about weak defaults and iterate through an unending list of possibilities.

Thinking that Eliezer thinks he knows that to any degree of specificity feels like I'm being weakmanned!

[Ngo][12:28]

I'm not claiming you have any specific beliefs

[Yudkowsky][12:29]

I suppose I have skepticism when other people dream up elaborately positive and beneficial reactions apparently drawn from some alternate nicer political universe that had an absolutely different response to Covid-19, and so on.

[Ngo][12:29]

But I'd guess that your models rule out, for instance, the US and China deeply cooperating on AI before it's caused any disasters

[Yudkowsky][12:30]

"Deeply"? Sure. That sounds like something that has never happened, and I'm generically skeptical about political things that go better than any political thing has ever gone before.

I guess we could talk about that? It doesn't seem like the most productive area, but maybe it lies upstream of more technical disagreements because we disagree about what AGI would actually have to do to have the world not end.

[Soares][12:31]

Cool. I claim it's time for a break, and then I nominate a little Eliezer gov't-response-overview followed by specific maybe-consequentialism-based-worries-aren't-a-problem-in-practice ideas from Richard.

[Ngo: ]

[Yudkowsky][12:32]

See you in 28mins

5.3. Epistemology, and assessing the idea of expected utility

[Ngo][13:03]

Some of my current thoughts are a reiteration of old despair: It feels to me like the typical Other within EA has no experience with discovering unexpected order, with operating a generalization that

you can expect will cover new cases even when that isn't immediately obvious, with operating that generalization to cover those new cases correctly, with seeing simple structures that generalize a lot and having that be a real and useful and technical experience; instead of somebody blathering in a non-expectation-constraining way about how "capitalism is responsible for everything wrong with the world", and being able to extend that to lots of cases.

I could try to use much simpler language in hopes that people actually look-at-the-water Feynman-style, like "navigating a path through time" instead of Consequentialism which is itself a step down from Expected Utility.

But you actually do lose something when you throw away the more technical concept. And then people still think that either you instantly see in the first second how something is a case of "navigating a path through time", or that this is something that people only do explicitly when visualizing paths through time using that mental terminology; or, if Eliezer says that it's "navigating time" anyways, this must be an instance of Eliezer doing that thing other people do when they talk about how "Capitalism is responsible for all the problems of the world". They have no experience operating genuinely useful, genuinely deep generalizations that extend to nonobvious things.

And in fact, being able to operate some generalizations like that is a lot of how I know what I know, in reality and in terms of the original knowledge that came before trying to argue that knowledge with people. So trying to convey the real source of the knowledge feels doomed. It's a kind of idea that our civilization has lost, like that college class Feynman ran into.

Ooops, didn't see this comment earlier. With respect to discovering unexpected order, one point that seems relevant is the extent to which that order provides predictive power. To what extent do you think that predictive successes in economics are important evidence for expected utility theory being a powerful formalism? (Or are there other ways in which it's predictively powerful that provide significant evidence?)

I'd be happy with a quick response to that, and then on geopolitics, here's a prompt to kick us off:

- If the only two actors involved in AGI development were the US and the UK governments, how much safer (or less safe) would you think we were compared with a world in which the two actors are the US and Chinese governments? How about a world in which the US government was a decade ahead of everyone else in reaching AGI?

[Yudkowsky][13:06]

I think that the Apollo space program is much deeper evidence for Utility. Observe, if you train protein blobs to run around the savanna, they also go to the moon!

If you think of "utility" as having something to do with the human discipline called "economics" then you are still thinking of it in a *much much much* more narrow way than I do.

[Ngo][13:07]

I'm not asking about evidence for utility as an abstraction in general, I'm asking for evidence based on successful predictions that have been made using it.

[Yudkowsky][13:10]

That doesn't tend to happen a lot, because all of the deep predictions that it makes are covered by shallow predictions that people made earlier.

Consider the following prediction of evolutionary psychology: Humans will enjoy activities associated with reproduction!

"What," says Simplicio, "you mean like dressing up for dates? I don't enjoy that part."

"No, you're overthinking it, we meant orgasms," says the evolutionary psychologist.

"But I already knew that, that's just common sense!" replies Simplicio.

"And yet it is very specifically a prediction of evolutionary psychology which is not made specifically by any other theory of human minds," replies the evolutionary psychologist.

"Not an advance prediction, just-so story, too obvious," replies Simplicio.

[Ngo][13:11]

Yepp, I agree that most of its predictions won't be new. Yet evolution is a sufficiently powerful theory that people have still come up with a range of novel predictions that derive from it.

Insofar as you're claiming that expected utility theory is also very powerful, then we should expect that it also provides some significant predictions.

[Yudkowsky][13:12]

An advance prediction of the notion of Utility, I suppose, is that if you train an AI which is otherwise a large blob of layers - though this may be inadvisable for other reasons - to the point where it starts solving lots of novel problems, that AI will tend to value aspects of outcomes with weights, and weight possible paths through time (the dynamic progress

of the environment), and use (by default, usually, roughly) the multiplication of these weights to allocate limited resources between mutually conflicting plans.

[Ngo][13:13]

Again, I'm asking for evidence in the form of successful predictions.

[Yudkowsky][13:14]

I predict that people will want some things more than others, think some possibilities are more likely than others, and prefer to do things that lead to stuff they want a lot through possibilities they think are very likely!

[Ngo][13:15]

It would be very strange to me if a theory which makes such strong claims about things we can't yet verify can't shed light on *anything* which we are in a position to verify.

[Yudkowsky][13:15]

If you think I'm deriving my predictions of catastrophic alignment failure through something *more exotic* than that, you're missing the reason *why I'm so worried*. It doesn't take intricate complicated exotic assumptions.

It makes the same kind of claims about things we can't verify yet as it makes about things we can verify right now.

[Ngo][13:16]

But that's very easy to do! Any theory can do that.

[Yudkowsky][13:17]

For example, if somebody wants money, and you set up a regulation which prevents them from making money, it predicts that the person will look for a new way to make money that bypasses the regulation.

[Ngo][13:17]

And yes, of course fitting previous data is important evidence in favour of a theory

[Yudkowsky][13:17]

[But that's very easy to do! Any theory can do that.]

False! Any theory can do that in the hands of a fallible agent which invalidly, incorrectly derives predictions from the theory.

[Ngo][13:18]

Well, indeed. But the very point at hand is whether the predictions you base on this theory are correctly or incorrectly derived.

[Yudkowsky][13:18]

It is not the case that every theory does an equally good job of predicting the past, given valid derivations of predictions.

Well, hence the analogy to evolutionary psychology. If somebody doesn't see the blatant obviousness of how sexual orgasms are a prediction specifically of evolutionary theory, because it's "common sense" and "not an advance prediction", what are you going to do? We can, in this case, with a *lot* more work, derive more detailed advance predictions about degrees of wanting that correlate in detail with detailed fitness benefits. But that's not going to convince anybody who overlooked the really blatant and obvious primary evidence.

What they're missing there is a sense of counterfactuals, of how the universe could just as easily have looked if the evolutionary origins of psychology were false: why should organisms want things associated with reproduction, why not instead have organisms running around that want things associated with rolling down hills?

Similarly, if optimizing complicated processes for outcomes hard enough, didn't produce cognitive processes that internally mapped paths through time and chose actions conditional on predicted outcomes, human beings would... not think like that? What am I supposed to say here?

[Ngo][13:24]

Let me put it this way. There are certain traps that, historically, humans have been very liable to fall into. For example, seeing a theory, which seems to match so beautifully and elegantly the data which we've collected so far, it's very easy to dramatically overestimate how much that data favours that theory. Fortunately, science has a very powerful social technology for avoiding this (i.e. making falsifiable predictions) which seems like approximately the only reliable way to avoid it - and yet you don't seem concerned at all about the lack of application of this technology to expected utility theory.

[Yudkowsky][13:25]

This is territory I covered in the Sequences, exactly because "well it didn't make a good enough advance prediction yet!" is an excuse that people use to reject evolutionary psychology, some other stuff I covered in the Sequences, and some very predictable lethaliies of AGI.

[Ngo][13:26]

With regards to evolutionary psychology: yes, there are some blatantly obvious ways in which it helps explain the data available to us. But there are also many people who have misapplied or overapplied evolutionary psychology, and it's very difficult to judge whether they have or have not done so, without asking them to make advance predictions.

[Yudkowsky][13:26]

I talked about the downsides of allowing humans to reason like that, the upsides, the underlying theoretical laws of epistemology (which are clear about why agents that reason validly or just unbiasedly would do that without the slightest hiccup), etc etc.

In the case of the theory "people want stuff relatively strongly, predict stuff relatively strongly, and combine the strengths to choose", what kind of advance prediction that no other theory could possibly make, do you expect that theory to make?

In the worlds where that theory is true, how should it be able to prove itself to you?

[Ngo][13:28]

I expect deeper theories to make more and stronger predictions.

I'm currently pretty uncertain if expected utility theory is a deep or shallow theory.

But deep theories tend to shed light in all sorts of unexpected places.

[Yudkowsky][13:30]

The fact is, when it comes to AGI (general optimization processes), we have only two major datapoints in our dataset, natural selection and humans. So you can either try to reason validly about what theories predict about natural selection and humans, even though we've already seen the effects of those; or you can claim to give up in great humble [modesty](#) while actually using other implicit theories instead to make all your predictions and be confident in them.

[Ngo][13:30]

I talked about the downsides of allowing humans to reason like that, the upsides, the underlying theoretical laws of epistemology (which are clear about why agents that reason validly or just unbiasedly would do that without the slightest hiccup), etc etc.

I'm familiar with your writings on this, which is why I find myself surprised here. I could understand a perspective of "yes, it's unfortunate that there are no advanced predictions, it's a significant weakness, I wish more people were doing this so we could better understand this vitally important theory". But that seems very different from your perspective here.

[Yudkowsky][13:32]

Oh, I'd love to be making predictions using a theory that made super detailed advance predictions made by no other theory which had all been borne out by detailed experimental observations! I'd also like ten billion dollars, a national government that believed everything I honestly told them about AGI, and a drug that raises IQ by 20 points.

[Ngo][13:32]

The very fact that we have only two major datapoints is exactly why it seems like such a major omission that a theory which purports to describe intelligent agency has not been used to make any successful predictions about the datapoints we do have.

[Yudkowsky][13:32][13:33]

This is making me think that you imagine the theory as something much more complicated and narrow than it is.

Just look at the water.

Not very special water with an index.

Just regular water.

People want stuff. They want some things more than others. When they do stuff they expect stuff to happen.

These are *predictions of the theory*. Not advance predictions, but predictions nonetheless.

[Ngo][13:33][13:33]

I'm accepting your premise that it's something deep and fundamental, and making the claim that deep, fundamental theories are likely to have a wide range of applications, including ones we hadn't previously thought of.

Do you disagree with that premise, in general?

[Yudkowsky][13:36]

I don't know what you really mean by "deep fundamental theory" or "wide range of applications we hadn't previously thought of", especially when it comes to structures that are this simple. It sounds like you're still imagining something I mean by Expected Utility which is some narrow specific theory like a particular collection of gears that are appearing in lots of places.

Are numbers a deep fundamental theory?

Is addition a deep fundamental theory?

Is probability a deep fundamental theory?

Is the notion of the syntax-semantics correspondence in logic and the notion of a generally semantically valid reasoning step, a deep fundamental theory?

[Ngo][13:38]

Yes to the first three, all of which led to very successful novel predictions.

[Yudkowsky][13:38]

What's an example of a novel prediction made by the notion of probability?

[Ngo][13:38]

Most applications of the central limit theorem.

[Yudkowsky][13:39]

Then I should get to claim every kind of optimization algorithm which used expected utility, as a successful advance prediction of expected utility? Optimal stopping and all the rest? Seems cheap and indeed invalid to me, and not particularly germane to whether these things appear inside AGIs, but if that's what you want, then sure.

[Ngo][13:39]

These are *predictions of the theory*. Not advance predictions, but predictions nonetheless.

I agree that it is a prediction of the theory. And yet it's also the case that smarter people than either of us have been dramatically mistaken about how well theories fit previously-collected data. (Admittedly we have advantages which they didn't, like a better understanding of cognitive biases - but it seems like you're ignoring the possibility of those cognitive biases applying to us, which largely negates those advantages.)

[Yudkowsky][13:42]

I'm not ignoring it, just adjusting my confidence levels and proceeding, instead of getting stuck in an infinite epistemic trap of self-doubt.

I don't live in a world where you either have the kind of detailed advance experimental predictions that should convince the most skeptical scientist and render you immune to all criticism, or, alternatively, you are suddenly in a realm beyond the reach of all epistemic authority, and you ought to cuddle up into a ball and rely only on wordless intuitions and trying to put equal weight on good things happening and bad things happening.

I live in a world where I proceed with very strong confidence if I have a detailed formal theory that made detailed correct advance predictions, and otherwise go around saying, "well, it sure looks like X, but we can be on the lookout for a miracle too".

If this was a matter of thermodynamics, I wouldn't even be talking like this, and we wouldn't even be having this debate.

I'd just be saying, "Oh, that's a perpetual motion machine. You can't build one of those. Sorry." And that would be the end.

Meanwhile, political superforecasters go on making well-calibrated predictions about matters much murkier and more complicated than these, often without anything resembling a clearly articulated theory laid forth at length, let alone one that had made specific predictions even retrospectively. They just go do it instead of feeling helpless about it.

[Ngo][13:45]

Then I should get to claim every kind of optimization algorithm which used expected utility, as a successful advance prediction of expected utility? Optimal stopping and all the rest? Seems cheap and indeed invalid to me, and not particularly germane to whether these things appear inside AGIs, but if that's what you want, then sure.

These seem better than nothing, but still fairly unsatisfying, insofar as I think they are related to more shallow properties of the theory.

Hmm, I think you're mischaracterising my position. I nowhere advocated for feeling helpless or curling up in a ball. I was just noting that this is a particularly large warning sign which has often been valuable in the past,

and it seemed like you were not only speeding past it blithely, but also denying the existence of this category of warning signs.

[Yudkowsky][13:48]

I think you're looking for some particular kind of public obeisance that I don't bother to perform internally because I'd consider it a wasted motion. If I'm lost in a forest I don't bother going around loudly talking about how I need a forest theory that makes detailed advance experimental predictions in controlled experiments, but, alas, I don't have one, so now I should be very humble. I try to figure out which way is north.

When I have a guess at a northerly direction, it would then be an error to proceed with as much confidence as if I'd had a detailed map and had located myself upon it.

[Ngo][13:49]

Insofar as I think we're less lost than you do, then the weaknesses of whichever forest theory implies that we're lost are relevant for this discussion.

[Yudkowsky][13:49]

The obeisance I make in that direction is visible in such statements as, "But this, of course, is a prediction about the future, which is well-known to be quite difficult to predict, in fact."

If my statements had been matters of thermodynamics and particle masses, I would *not* be adding that disclaimer.

But most of life is not a statement about particle masses. I have some idea of how to handle that. I do not need to constantly recite disclaimers to myself about it.

I know how to proceed when I have only a handful of data points which have already been observed and my theories of them are retrospective theories. This happens to me on a daily basis, eg when dealing with human beings.

[Soares][13:50]

(I have a bit of a sense that we're going in a circle. It also seems to me like there's some talking-past happening.)

(I suggest a 5min break, followed by EY attempting to paraphrase RN to his satisfaction and vice versa.)

[Yudkowsky][13:51]

I'd have more trouble than usual paraphrasing RN because epistemic helplessness is something I find painful to type out.

[Soares][13:51]

(I'm also happy to attempt to paraphrase each point as I see it; it may be that this smooths over some conversational wrinkle.)

[Ngo][13:52]

Seems like a good suggestion. I'm also happy to move on to the next topic. This was meant to be a quick clarification.

[Soares][13:52]

nod. It does seem to me like it possibly contains a decently sized meta-crux, about what sorts of conclusions one is licensed to draw from what sorts of observations

that, eg, might be causing Eliezer's probabilities to concentrate but not Richard's.

[Yudkowsky][13:52]

Yeah, this is in the opposite direction of "more specificity".

[Soares: 😊] [Ngo: 😊]

I frankly think that most EAs suck at explicit epistemology, OpenPhil and FHI affiliated EAs are not much of an exception to this, and I expect I will have more luck talking people out of specific errors than talking them out of the infinite pit of humble ignorance considered abstractly.

[Soares][13:54]

Ok, that seems to me like a light bid to move to the next topic from both of you, my new proposal is that we take a 5min break and then move to the next topic, and perhaps I'll attempt to paraphrase each point here in my notes, and if there's any movement in the comments there we can maybe come back to it later.

[Ngo: 🤞]

[Ngo][13:54]

Broadly speaking I am also strongly against humble ignorance (albeit to a lesser extent than you are).

[Yudkowsky][13:55]

I'm off to take a 5-minute break, then!

5.4. Government response and economic impact

[Ngo][14:02]

A meta-level note: I suspect we're around the point of hitting significant diminishing marginal returns from this format. I'm open to putting more time into the debate (broadly construed) going forward, but would probably want to think a bit about potential changes in format.

[Soares][14:04, moved two up in log]

A meta-level note: I suspect we're around the point of hitting significant diminishing marginal returns from this format. I'm open to putting more time into the debate (broadly construed) going forward, but would probably want to think a bit about potential changes in format.

(Noted, thanks!)

[Yudkowsky][14:03]

I actually think that may just be a matter of at least one of us, including Nate, having to take on the thankless job of shutting down all digressions into abstractions and the meta-level.

[Ngo][14:05]

I actually think that may just be a matter of at least one of us, including Nate, having to take on the thankless job of shutting down all digressions into abstractions and the meta-level.

I'm not so sure about this, because it seems like some of the abstractions are doing a lot of work.

[Yudkowsky][14:03][14:04]

Anyways, government reactions?

It seems to me like the best observed case for government reactions - which I suspect is no longer available in the present era as a possibility - was the degree of cooperation between the USA and Soviet Union about avoiding nuclear exchanges.

This included such incredibly extravagant acts of cooperation as installing a direct line between the President and Premier!

which is not what I would really characterize as very "deep" cooperation, but it's more than a lot of cooperation you see nowadays.

More to the point, both the USA and Soviet Union proactively avoided doing anything that might lead towards starting down a path that led to a full nuclear exchange.

[Ngo][14:04]

The question I asked earlier:

- If the only two actors involved in AGI development were the US and the UK governments, how much safer (or less safe) would you think we were compared with a world in which the two actors are the US and Chinese governments? How about a world in which the US government was a decade ahead of everyone else in reaching AGI?

[Yudkowsky][14:05]

They still provoked one another a lot, but, whenever they did so, tried to do so in a way that wouldn't lead to a full nuclear exchange.

It was mutually understood to be a strategic priority and lots of people on both sides thought a lot about how to avoid it.

I don't know if that degree of cooperation ever got to the fantastic point of having people from *both* sides in the *same* room brainstorming *together* about how to avoid a full nuclear exchange, because that is, like, more cooperation than you would normally expect from two governments, but it wouldn't *shock* me to learn that this had ever happened.

It seems obvious to me that if some situation developed nowadays which increased the profile possibility of a nuclear exchange between the USA and Russia, we would not currently be able to do anything like installing a Hot Line between the US and Russian offices if such a Hot Line had not already been installed. This is lost social technology from a lost golden

age. But still, it's not unreasonable to take this as the upper bound of attainable cooperation; it's been observed within the last 100 years.

Another guess for how governments react is a very simple and robust one backed up by a huge number of observations:

They don't.

They have the same kind of advance preparation and coordination around AGI, in advance of anybody getting killed, as governments had around the mortgage crisis of 2007 in advance of any mortgages defaulting.

I am not sure I'd put this probability over 50% but it's certainly by far the largest probability over any competitor possibility specified to an equally low amount of detail.

I would expect anyone whose primary experience was with government, who was just approaching this matter and hadn't been talked around to weird exotic views, to tell you the same thing as a matter of course.

[Ngo][14:10]

But still, it's not unreasonable to take this as the upper bound of attainable cooperation; it's been observed within the last 100 years.

Is this also your upper bound conditional on a world that has experienced a century's worth of changes within a decade, and in which people are an order of magnitude wealthier than they currently are?

I am not sure I'd put this probability over 50% but it's certainly by far the largest probability over any competitor possibility specified to an equally low amount of detail.

which one was this? US/UK?

[Yudkowsky][14:12][14:14]

Assuming governments do react, we have the problem of "What kind of heuristic could have correctly led us to forecast that the US's reaction to a major pandemic would be for the FDA to ban hospitals from doing in-house Covid tests? What kind of mental process could have led us to make that call?" And we couldn't have gotten it exactly right, because the future is hard to predict; the best heuristic I've come up with, that feels like it at least would not have been *surprised* by what actually happened, is, "The government will react with a flabbergasting level of incompetence, doing exactly the wrong thing, in some unpredictable specific way."

which one was this? US/UK?

I think if we're talking about any single specific government like the US or UK then the probability is over 50% that they don't react in any advance

coordinated way to the AGI crisis, *to a greater and more effective degree* than they "reacted in an advance coordinated way" to pandemics before 2020 or mortgage defaults before 2007.

Maybe *some* two governments somewhere on Earth will have a high-level discussion between two cabinet officials.

[Ngo][14:14]

That's one lesson you could take away. Another might be: governments will be very willing to restrict the use of novel technologies, even at colossal expense, in the face of even a small risk of large harms.

[Yudkowsky][14:15]

That's one lesson you could take away. Another might be: governments will be very willing to restrict the use of novel technologies, even at colossal expense, in the face of even a small risk of large harms.

I just... don't know what to do when people talk like this.

It's so absurdly, absurdly optimistic.

It's taking a massive failure and trying to find exactly the right abstract gloss to put on it that makes it sound like exactly the right perfect thing will be done next time.

This just - isn't how to understand reality.

This isn't how superforecasters think.

This isn't sane.

[Soares][14:16]

(be careful about ad hominem)

(Richard might not be doing the insane thing you're imagining, to generate that sentence, etc)

[Ngo][14:17]

Right, I'm not endorsing this as my mainline prediction about what happens. Mainly what I'm doing here is highlighting that your view seems like one which cherrypicks *pessimistic* interpretations.

[Yudkowsky][14:18]

That abstract description "governments will be very willing to restrict the use of novel technologies, even at colossal expense, in the face of even a small risk of large harms" does not in fact apply very well to the FDA banning hospitals from using their well-established in-house virus tests, at risk of the alleged harm of some tests giving bad results, when in fact the CDC's tests were giving bad results and much larger harms were on the way because of bottlenecked testing; and that abstract description should have applied to an effective and globally coordinated ban against gain-of-function research, which *didn't* happen.

[Ngo][14:19]

Alternatively: what could have led us to forecast that many countries will impose unprecedentedly severe lockdowns.

[Yudkowsky][14:19][14:21][14:21]

Well, I didn't! I didn't even realize that was an option! I thought Covid was just going to rip through everything.

(Which, to be clear, it still may, and Delta arguably is in the more primitive tribal areas of the USA, as well as many other countries around the world that can't afford vaccines financially rather than epistemically.)

But there's a really really basic lesson here about the different style of "sentences found in political history books" rather than "sentences produced by people imagining ways future politics could handle an issue successfully".

Reality is *so much worse* than people imagining what might happen to handle an issue successfully.

[Ngo][14:21][14:21][14:22]

I might nudge us away from covid here, and towards the questions I asked before.

The question I asked earlier:

- If the only two actors involved in AGI development were the US and the UK governments, how much safer (or less safe) would you think we were compared with a world in which the two actors are the US and Chinese governments? How about a world in which the US government was a decade ahead of everyone else in reaching AGI?

This being one.

"But still, it's not unreasonable to take this as the upper bound of

attainable cooperation; it's been observed within the last 100 years." Is this also your upper bound conditional on a world that has experienced a century's worth of changes within a decade, and in which people are an order of magnitude wealthier than they currently are?

And this being the other.

[Yudkowsky][14:22]

Is this also your upper bound conditional on a world that has experienced a century's worth of changes within a decade, and in which people are an order of magnitude wealthier than they currently are?

I don't expect this to happen at all, or even come remotely close to happening; I expect AGI to kill everyone before self-driving cars are commercialized.

[Yudkowsky][16:29] (Nov. 14 follow-up comment)

(This was incautiously put; maybe strike "expect" and put in "would not be the least bit surprised if" or "would very tentatively guess that".)

[Ngo][14:23]

ah, I see

Okay, maybe here's a different angle which I should have been using. What's the most impressive technology you expect to be commercialised before AGI kills everyone?

[Yudkowsky][14:24]

If the only two actors involved in AGI development were the US and the UK governments, how much safer (or less safe) would you think we were compared with a world in which the two actors are the US and Chinese governments?

Very hard to say; the UK is friendlier but less grown-up. We would obviously be VASTLY safer in any world where only two centralized actors (two effective decision processes) could ever possibly build AGI, though not safe / out of the woods / at over 50% survival probability.

How about a world in which the US government was a decade ahead of everyone else in reaching AGI?

Vastly safer and likewise impossibly miraculous, though again, not out of the woods at all / not close to 50% survival probability.

What's the most impressive technology you expect to be commercialised before AGI kills everyone?

This is incredibly hard to predict. If I actually had to predict this for some reason I would probably talk to Gwern and Carl Shulman. In principle, there's nothing preventing me from knowing something about Go which lets me predict in 2014 that Go will probably fall in two years, but in practice I did not do that and I don't recall anybody else doing it either. It's really quite hard to figure out how much cognitive work a domain requires and how much work known AI technologies can scale to with more compute, let alone predict AI breakthroughs.

[Ngo][14:27]

I'd be happy with some very rough guesses

[Yudkowsky][14:27]

If you want me to spin a scifi scenario, I would not be surprised to find online anime companions carrying on impressively humanlike conversations, because this is a kind of technology that can be deployed without major corporations signing on or regulatory approval.

[Ngo][14:28]

Okay, this is surprising; I expected something more advanced.

[Yudkowsky][14:29]

Arguably AlphaFold 2 is already more advanced than that, along certain dimensions, but it's no coincidence that afaik people haven't really done much with AlphaFold 2 and it's made no visible impact on GDP.

I expect GDP not to depart from previous trendlines before the world ends, would be a more general way of putting it.

[Ngo][14:29]

What's the ~~most~~ least impressive technology that your model strongly rules out happening before AGI kills us all?

[Yudkowsky][14:30]

you mean least impressive?

[Ngo][14:30]

oops, yes

That seems like a structurally easier question to answer

[Yudkowsky][14:30]

"Most impressive" is trivial. "Dyson Spheres" answers it.

Or, for that matter, "perpetual motion machines".

[Ngo][14:31]

Ah yes, I was thinking that Dyson spheres were a bit too prosaic

[Yudkowsky][14:32]

My model mainly rules out that we get to certain points and then hang around there for 10 years while the technology gets perfected, commercialized, approved, adopted, ubiquitized enough to produce a visible trendline departure on the GDP graph; not so much various technologies themselves being initially demonstrated in a lab.

I expect that the people who build AGI can build a self-driving car if they want to. Getting it approved and deployed before the world ends is quite another matter.

[Ngo][14:33]

OpenAI has commercialised GPT-3

[Yudkowsky][14:33]

Hasn't produced much of a bump in GDP as yet.

[Ngo][14:33]

I wasn't asking about that, though

I'm more interested in judging how hard you think it is for AIs to take over the world

[Yudkowsky][14:34]

I note that it seems to me like there is definitely a kind of thinking here, which, if told about GPT-3 five years ago, would talk in very serious tones about how much this technology ought to be predicted to shift GDP, and whether we could bet on that.

By "take over the world" do you mean "turn the world into paperclips" or "produce 10% excess of world GDP over predicted trendlines"?

[Ngo][14:35]

Turn world into paperclips

[Yudkowsky][14:36]

I expect this mainly happens as a result of superintelligence, which is way up in the stratosphere far above the minimum required cognitive capacities to get the job done?

The interesting question is about humans trying to deploy a corrigible AGI thinking in a restricted domain, trying to flip the gameboard / "take over the world" without full superintelligence?

I'm actually not sure what you're trying to get at here.

[Soares][14:37]

(my guess, for the record, is that the crux Richard is attempting to drive for here, is centered more around something like "will humanity spend a bunch of time in the regime where there are systems capable of dramatically increasing world GDP, and if not how can you be confident of that from here")

[Yudkowsky][14:38]

This is not the sort of thing I feel Confident about.

[Yudkowsky][16:31] (Nov. 14 follow-up comment)

(My confidence here seems understated. I am very pleasantly surprised if we spend 5 years hanging around with systems that can dramatically increase world GDP and those systems are actually being used for that. There isn't one dramatic principle which prohibits that, so I'm not Confident, but it requires multiple nondramatic events to go not as I expect.)

[Ngo][14:38]

Yeah, that's roughly what I'm going for. Or another way of putting it: we have some disagreements about the likelihood of humans being able to get an AI to do a pivotal act which saves the world. So I'm trying to get some estimates for what the hardest act you think humans *can* get an AI to do is.

[Soares][14:39]

(and that a difference here causes, eg, Richard to suspect the relevant geopolitics happen after a century of progress in 10y, everyone being suddenly much richer in real terms, and a couple of warning shots, whereas Eliezer expects the relevant geopolitics to happen the day after tomorrow, with "realistic human-esque convos" being the sort of thing we get instead of warning shots)

[Ngo: ]

[Yudkowsky][14:40]

I mostly do not expect pseudo-powerful but non-scalable AI powerful enough to increase GDP, hanging around for a while. But if it happens then I don't feel I get to yell "what happened?" at reality, because there's an obvious avenue for it to happen: something GDP-increasing proved tractable to non-deeply-general AI systems.

where GPT-3 is "not deeply general"

[Ngo][14:40]

Again, I didn't ask about GDP increases, I asked about impressive acts (in order to separate out the effects of AI capabilities from regulatory effects, people-having-AI-but-not-using-it, etc).

Where you can use whatever metric of impressiveness you think is reasonable.

[Yudkowsky][14:42]

so there's two questions here, one of which is something like, "what is the most impressive thing you can do while still being able to align stuff and make it corrigible", and one of which is "if there's an incorrigible AI whose deeds are being exhibited by fools, what impressive things might it do short of ending the world".

and these are both problems that are hard for the same reason I did not predict in 2014 that Go would fall in 2016; it can in fact be quite hard - even with a domain as fully lawful and known as Go - to figure out which problems will fall to which level of cognitive capacity.

[Soares][14:43]

Nate's attempted rephrasing: EY's model might not be confident that there's not big GDP boosts, but it does seem pretty confident that there isn't some "half-capable" window between the shallow-pattern-memorizer

stuff and the scary-laserlike-consequentialist stuff, and in particular Eliezer seems confident humanity won't slowly traverse that capability regime

[Yudkowsky][14:43]

that's... allowed? I don't get to yell at reality if that happens?

[Soares][14:44]

and (shakier extrapolation), that regime is where a bunch of Richard's hope lies (eg, in the beginning of that regime we get to learn how to do practical alignment, and also the world can perhaps be saved midway through that regime using non-laserlike-systems)

[Ngo: ]

[Yudkowsky][14:45]

so here's an example of a thing I don't think you can do without the world ending: get an AI to build a nanosystem or biosystem which can synthesize two strawberries identical down to the cellular but not molecular level, and put them on a plate

this is why I use this capability as the definition of a "powerful AI" when I talk about "powerful AIs" being hard to align, if I don't want to start by explicitly arguing about pivotal acts

this, I think, is going to end up being first doable using a laserlike world-ending system

so even if there's a way to do it with no lasers, that happens later and the world ends before then

[Ngo][14:47]

Okay, that's useful.

[Yudkowsky][14:48]

it feels like the critical bar there is something like "invent a whole engineering discipline over a domain where you can't run lots of cheap simulations in full detail"

[Ngo][14:49]

(Meta note: let's wrap up in 10 mins? I'm starting to feel a bit sleepy.)

[Yudkowsky:] [Soares:]

This seems like a pretty reasonable bar

Let me think a bit about where to go from that

While I'm doing so, since this question of takeoff speeds seems like an important one, I'm wondering if you could gesture at your biggest disagreement with this post: <https://sideways-view.com/2018/02/24/takeoff-speeds/>

[Yudkowsky][14:51]

Oh, also in terms of scifi possibilities, I can imagine seeing 5% GDP loss because text transformers successfully scaled to automatically filing lawsuits and environmental impact objections.

My read on the entire modern world is that GDP is primarily constrained by bureaucratic sclerosis rather than by where the technological frontiers lie, so AI ends up impacting GDP mainly insofar as it allows new ways to bypass regulatory constraints, rather than insofar as it allows new technological capabilities. I expect a sudden transition to paperclips, not just because of how fast I expect cognitive capacities to scale over time, but because nanomachines eating the biosphere bypass regulatory constraints, whereas earlier phases of AI will not be advantaged relative to all the other things we have the technological capacity to do but which aren't legal to do.

[Shah][12:13] (Sep. 21 follow-up comment)

My read on the entire modern world is that GDP is primarily constrained by bureaucratic sclerosis rather than by where the technological frontiers lie

This is a fair point and updates me somewhat towards fast takeoff as operationalized by Paul, though I'm not sure how much it updates me on p(doom).

Er, wait, really fast takeoff as operationalized by Paul makes less sense as a thing to be looking for -- presumably we die before any 1 year doubling. Whatever, it updates me somewhat towards "less deployed stuff before scary stuff is around"

[Ngo][14:56]

Ah, interesting. What are the two or three main things in that category?

[Yudkowsky][14:57]

mRNA vaccines, building houses, building cities? Not sure what you mean there.

[Ngo][14:57]

"things we have the technological capacity to do but which aren't legal to do"

[Yudkowsky][14:58][15:00]

Eg, you might imagine, "What if AIs were smart enough to build houses, wouldn't that raise GDP?" and the answer is that we already have the pure technology to manufacture homes cheaply, but the upright-stick-construction industry already successfully lobbied to get it banned as it was starting to develop, by adding on various constraints; so the question is not "Is AI advantaged in doing this?" but "Is AI advantaged at bypassing regulatory constraints on doing this?" Not to mention all the other ways that building a house in an existing city is illegal, or that it's been made difficult to start a new city, etcetera.

"What if AIs could design a new vaccine in a day?" We can already do that. It's no longer the relevant constraint. Bureaucracy is the process-limiting constraint.

I would - looking in again at the Sideways View essay on takeoff speeds - wonder whether it occurred to you, Richard, to ask about what detailed predictions all the theories there had made.

After all, a lot of it is spending time explaining why the theories there *shouldn't* be expected to retrodict even the data points we *have* about progress rates over hominid evolution.

Surely you, being the evenhanded judge that you are, must have been reading through that document saying, "My goodness, this is even worse than retrodicting a few data points!"

A lot of why I have a bad taste in my mouth about certain classes of epistemological criticism is my sense that certain sentences tend to be uttered on *incredibly* selective occasions.

[Ngo][14:59][15:06]

Some meta thoughts: I now feel like I have a pretty reasonable broad outline of Eliezer's views. I haven't yet changed my mind much, but plausibly mostly because I haven't taken the time to internalise those views; once I ruminate on them a bunch, I expect my opinions will shift (uncertain how far; unlikely to be most of the way).

Meta thoughts (continued): Insofar as a strong disagreement remains after that (which it probably will) I feel pretty uncertain about what would resolve it. Best guess is that I should write up some longer essays that try to tie a bunch of disparate strands together.

Near the end it seemed like the crux, to a surprising extent, hinged on this question of takeoff speeds. So the other thing which seems like it'd plausibly help a lot is Eliezer writing up a longer version of his response to Paul's Takeoff Speeds post.

(Just as a brief comment, I don't find the "bureaucratic sclerosis" explanation very compelling. I do agree that regulatory barriers are a huge problem, but they still don't seem nearly severe enough to cause a fast takeoff. I don't have strong arguments for that position right now though.)

[Soares][15:12]

This seems like a fine point to call it!

Some wrap-up notes

- I had the impression this round was a bit more frustrating than last rounds. Thanks all for sticking with things 😊
- I have a sense that Richard was making a couple points that didn't quite land. I plan to attempt to articulate versions of them myself in the interim.
- Richard noted he had a sense we're in decreasing return territory. My own sense is that it's worth having at least one more discussion in this format about specific non-consequentialist plans Richard may have hope in, but I also think we shouldn't plow forward in spite of things feeling less useful, and I'm open to various alternative proposals.

In particular, it seems maybe plausible to me we should have a pause for some offline write-ups, such as Richard digesting a bit and then writing up some of his current state, and/or Eliezer writing up some object-level response to the takeoff speed post above?

[Ngo: 👍]

(I also could plausibly give that a go myself, either from my own models or from my model of Eliezer's model which he could then correct)

[Ngo][15:15]

Thanks Nate!

I endorse the idea of offline writeups

[Soares][15:17]

Cool. Then I claim we are adjourned for the day, and Richard has the ball on digesting & doing a write-up from his end, and I have the ball on both writing up my attempts to articulate some points, and on either Eliezer or I writing some takes on timelines or something.

(And we can coordinate our next discussion, if any, via email, once the write-ups are in shape.)

[Yudkowsky][15:18]

I also have a sense that there's more to be said about specifics of govt stuff or specifics of "ways to bypass consequentialism" and that I wish we could spend at least one session trying to stick to concrete details only

Even if it's not where cruxes ultimately lie, often you learn more about the abstract by talking about the concrete than by talking about the abstract.

[Soares][15:22]

(I, too, would be enthusiastic to see such a discussion, and Richard, if you find yourself feeling enthusiastic or at least not-despairing about it, I'd happily moderate.)

[Yudkowsky][15:37]

(I'm a little surprised about how poorly I did at staying concrete after saying that aloud, and would nominate Nate to take on the stern duty of blowing the whistle at myself or at both of us.)

App and book recommendations for people who want to be happier and more productive

If somebody asks you the same question more than ten times, that's probably a sign it's time to write a blog post about it. So after being asked by far more than ten people about what books and resources I recommend, both for EAs more broadly and for people interested in charity entrepreneurship more specifically, here it is! After spending roughly the last 15 years optimizing nearly constantly, these are the systems and products I recommend.

If you know better versions of what I recommend, please share it in the comments! While I recommend the apps I use, I'm sure there's ones out there that do the same thing but in a better way and I just haven't had the time or energy to pay the switching costs yet.

Apps and Extensions

- **Game-changers.** Everybody should be using these.
 - [Video Speed Controller](#). This allows you to hot key change the speed of videos anywhere on Chrome. It also doesn't limit you to 2x, which so many apps do for some reason. You'll never (involuntarily) watch things on 1x again.
 - **Clipboard history** with [CopyQ](#) (Mac) or [here](#) for Windows. Absolute game changer. It remembers everything you've copy-pasted and you can click it from a list or use shortcut keys to paste them again in the future. Saves you so much time and hassle. It's hard to describe how much this changes how you use your computer.
 - [Switch between your two most recent tabs](#). Use the shortcut Alt + Q to switch between your two most recently used tabs. It's like alt-tab but for tabs instead of windows. I can't imagine navigating a computer without this. It feels crippling. I know that there are better ones that allow you cycle through multiple tabs, not just your most recent. If you know of one, recommend it in the comments! I just haven't had the spare time to optimize this more.
 - [Google docs quick create](#). Shortcut key or single click to automatically create a new google document or spreadsheet. Saves a *ton* of time.
- **Very good.** Will be extremely useful for a lot of, but not all, people
 - [Quickcompose](#). You know how easy it is to get distracted by your inbox when you need to send an email? Quick compose makes it so that you can open up a window that's just a compose window so you can't get distracted by new emails.
 - [I don't care about cookies](#). Makes it so that Chrome automatically accepts cookies, so you never have to click "I accept cookies" ever again.
 - [StayFocusd](#). Limits your time on social media.
 - [News Feed Eradicator](#). Eradicates news feeds from Twitter and Facebook and probably other social media as well. Saves you innumerable hours a year.

- **[SmileAlways](#)**. Makes it so that it automatically always opens Amazon Smile so that all your purchases contribute to charity. AMF is an option there. Nice passive impact opportunity.
- **[Beeminder](#)**. You only pay if you don't achieve the goals you set for yourself. [Stickk](#) is similar and allows you to make the money to go to an anti-charity of your choice and easily add an accountability buddy. I like the UI of Beeminder better since it's better for tracking my progress, but tastes will vary.
- **[Focusmate](#)**. Have an immediate, remote, videochat accountability buddy to help you work on things for 30 or 60 minute increments. Really good for getting done those things you don't usually want to do or if you're having trouble focusing (e.g. paperwork, homework, etc).
- **[Keysmith](#)**. Allows you to make a "program" by having the computer watch what you do and try to mimic your actions (e.g. open a window, type X, press these keys, etc). Allows you to code with no coding. It's not perfect, but it works for a lot of things. I use it to automate really repetitive tasks. (Mac only. Probably a Windows version somewhere.)
- **[Habitica](#)**. Gamified habit formation with a social accountability mechanism I find very compelling.
- **[Emojis everywhere](#)**. The windows key + period(.) opens the search bar for emojis on *any app* on your computer. It's ctrl + command + space for Mac. This can make you an emoji boss. ☺
- **Turn everything into audio**. This is a super valuable and useful.
 - **[Nonlinear Library](#)** of course! This has been a game-changer for me. I always wanted to read more of the Forum but never found the time and motivation. Since it's so easy, I now read the best articles every day. I just listen to it in the morning while getting ready or while commuting to work. But there's no customizability or ability to add your own idiosyncratic reading list, which means it's incomplete. For the rest, I use:
 - **[NaturalReader](#)** for articles on your computer
 - **[Evie](#)** for books on phone
 - **[@Voice](#)** for articles on phone
- **[Pocket](#)** and **[Save to Pocket](#)** chrome extension to quickly send an article to your Pocket for later
- **[Workflowy](#)**. I've also heard good things about [Roam](#) but seems similar enough to Workflowy to not be worth the switching costs for me. I'd play around with both to figure out which you prefer before committing. Game changer for thinking better. I use it for [steelman solitaire](#), a technique I developed that is great for thinking things through really deeply.
- **[Messenger](#)**. Some people message people by actually opening Facebook. This is a recipe for distraction and wasting your time on social media. Always use the messenger app ([desktop](#) or [phone](#)) to talk to people on Facebook.
- **All IMs on desktop**. Some people only do WhatsApp, texting, or other forms of IM on their phone. This is crazy. No matter how fast you type on a phone, you're slow compared to a computer. Almost all of these apps have the ability to do them from your computer. Here's instructions for [WhatsApp](#) and [Android](#) texting.
- **[Boomerang for Gmail](#)**. Sends back your email if you haven't heard a response in X period of time. Perfect for being able to follow up with people at the right time, especially weeks to months later. Makes you feel like you're the most organized person on the planet.

- **A password manager.** It saves you so much time filling in forms, which it'll auto-fill for you. Far better than existing autofills on Chrome. I use [Dashlane](#) which is also a VPN.
- **New Tab Motivational Quotes.** Chrome extension that makes every new tab show a motivational quote. If motivational quotes work for you, this can often help you stay energized throughout the day.
- **CopyAllUrls.** Copies the URLs of all the tabs you have open. Comes in handy a lot.

Websites

- **Libgen.** It's sci-hub but for books. And for scientific articles and comics too! [Here's a guide on how to use it.](#) It's pretty easy to use and has pretty much all books. If it's not available in your country, use a VPN.
- **Pinterest for motivation.** Pinterest allows you to train your own ML algorithms for particular purposes by making "boards". I have an inspiration board where I just add anything I find inspiring. Now it's trained so well on my tastes that it gives me eerily good recommendations. I've saved many a work day because of this.

Books and Blogs

- **How To Measure Anything.** Classic in the EA movement for a reason. Lots of practical techniques.
- **Copyblogger.** It's the writing class written by people who actually write for a living in the real world (unlike your English teachers). Absolutely essential for comms if you're running an organization or want to write for a cause. And, since they're good at writing, it's a really easy read of course. Here's [some very unpolished notes I wrote while reading it.](#) [Their book on headlines](#) is probably the highest value thing to read of theirs. You have to give them your email address to get the free e-book, but it's totally worth it. They also have a ton of free content there that I highly recommend.
- **Lean Startup.** Enough said about this already.
- **Atomic Habits.** The best habit development book out there. See also my [blog post here about my framework for onboarding habits.](#)
- **The Mom Test.** Survey design that they don't teach you in school. How to get actionable answers when you're a practitioner trying to actually learn for real work, not just a theoretician in an ivory tower. Short book with really good insights.
- **The 4-Hour Work Week.** Tim Ferris is probably the most instrumentally rational person I know of. The idea of passive income is incredibly high leverage, and can be cross-applied to altruism ("passive impact"). I think a lot more EAs should be pursuing passive incomes or early retirement ([this is a great blog on that topic](#)).
- **Actually doing the exercises in books.** This isn't a book recommendation per se, but I figured it belonged here. You really really *really* need to become the sort of person who actually does the exercises in the books. This is huge for you *actually* improving.

Physical products

- [**Improve your posture with a laptop stand.**](#) The easiest intervention ever. Costs \$30 and you will have way less back pain, better posture, and look better, and with no ongoing effort required. Here's a [twitter thread I wrote about posture and how it's one of the best "bang for your buck" interventions to improve your attractiveness.](#)
- [**Mouse with a million programmable buttons.**](#) If you love shortcut keys (which you totally should), you will love mice with programmable buttons. You can make it have over 30 shortcut keys at the slight move of your thumb.
- [**Kindle cover to make it look nice.**](#) Half of the reason why people read paper books is because of their feel. It just feels nicer. Fix this by getting a cover for your Kindle that makes it feel beautiful to you. E-readers are better for reading in almost every single dimension. Don't let aesthetics stop you from reading.
- [**Travel-friendly workout gear.**](#) These elastic bands allow you to travel with your "weights" but they weigh practically nothing since they're just light plastic.
- [**A little canister attached to your key chain**](#) so that you can never be without your favorite spicy pepper powder.
- [**Longlasting lipstick.**](#) Don't bother having to re-apply all the time. This is the best I've found so far.
- [**IUDs.**](#) IUDs last for ten years, so you don't have to pay attention to birth control again for literally a *decade*. Also, if you get the hormonal one, you have a decent odds of literally stopping your periods. Imagine not having to deal with all of that for a decade! I forget the numbers, but it's somewhere in the range of 10% of women stop having periods while also having the most guaranteed birth control of all alternatives. The cost and pain savings are incomprehensible.
- [**Better, less gross insect repellant.**](#) This stuff works just as well as DEET and it doesn't smell bad. It's more likely putting on pleasant sunscreen instead of squirting yourself with death-chemicals, which is my usual experience with DEET-based products.
- [**Erasable pens.**](#) Pens are clearly better than pencils in that you can write on more surfaces and have better colour selection. The only problem is you can't erase them. Unless they're erasable pens that is, then they strictly dominate. These are the best I've found that can erase well and write on the most surfaces.

Games

- [**Game of Trust.**](#) Ten minute introduction to game theory in a really memorable way that will help you deeply learn some of the basics.
- [**We Become What We Behold.**](#) Profound five minute game. Illustrates how the effects of paying attention to what's interesting causes polarization and strife.
- [**Democracy.**](#) A really complicated game where you try to run a country. It gives you a lot of empathy for the tough job of politicians and each policy you make shows a writeup explaining the arguments for and against. It definitely has a liberal slant, but they do a decent job of steelmanning both sides, which made me feel I learned a lot more than the usual political content out there.
- [**1979 Revolution.**](#) The best way to really feel what it's like to be part of a revolution. It's a documentary-game of the Iranian revolution in 1979 and being the one making the decisions adds a really interesting element to it.

- [**Florence**](#). The most beautiful and deeply happy piece of art I've experienced in the last decade. Perhaps the most realistic fictional portrayal of relationships I've ever seen. And it's a phone game! It uses the medium for storytelling in really creative, metaphorical, and beautiful ways. I can't recommend it enough. It's about a 30 minute game time and totally worth the \$3.

Soares, Tallinn, and Yudkowsky discuss AGI cognition

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

This is a collection of follow-up discussions in the wake of Richard Ngo and Eliezer Yudkowsky's [Sep. 5-8](#) and [Sep. 14](#) conversations.

Color key:

Chat Google Doc content Inline comments

7. Follow-ups to the Ngo/Yudkowsky conversation

[Bensinger][1:50] (Nov. 23 follow-up comment)

A general background note: Readers who aren't already familiar with ethical injunctions or the unilateralist's curse should probably read [Ends Don't Justify Means \(Among Humans\)](#), along with an explanation of [the unilateralist's curse](#).

7.1. Jaan Tallinn's commentary

[Tallinn][6:38] (Sep. 18)

thanks for the interesting debate! here are my comments so far: [GDocs link]

[Tallinn] (Sep. 18 Google Doc)

meta

a few meta notes first:

- i'm happy with the below comments being shared further without explicit permission – just make sure you respect the sharing constraints of the discussion that they're based on;
- there's a lot of content now in the debate that branches out in multiple directions – i suspect a strong distillation step is needed to make it coherent and publishable;
- the main purpose of this document is to give a datapoint how the debate is coming across to a reader – it's very probable that i've misunderstood some things, but that's the point;
- i'm also largely using my own terms/metaphors – for additional triangulation.

pit of generality

it feels to me like the main crux is about the topology of the space of cognitive systems in combination with what it implies about takeoff. here's the way i understand eliezer's position:

there's a "pit of generality" attractor in cognitive systems space: once an AI system gets sufficiently close to the edge ("past the atmospheric turbulence layer"), it's bound to improve in catastrophic manner;

[Yudkowsky][11:10] (Sep. 18 comment)

it's bound to improve in catastrophic manner

I think this is true with quite high probability about an AI that gets high *enough*, if not otherwise corrigibilized, boosting up to strong superintelligence - this is what it means metaphorically to get "past the atmospheric turbulence layer".

"High enough" should not be very far above the human level and *may* be below it; John von Neumann with the ability to run some chains of thought at high serial speed, access to his own source code, and the ability to try branches of himself, seems like he could very likely do this, possibly modulo his concerns about stomping his own utility function making him more cautious.

People noticeably less smart than von Neumann might be able to do it too.

An AI whose components are more modular than a human's and more locally testable might have an easier time of the whole thing; we can imagine the FOOM getting rolling from something that was in some sense dumber than human.

But the *strong* prediction is that when you get well above the von Neumann level, why, that is *clearly* enough, and things take over and go Foom. The lower you go from that threshold, the less sure I am that it counts as "out of the atmosphere". This epistemic humility on my part should not be confused for knowledge of a constraint on the territory that requires AI to go far above humans to Foom. Just as DL-based AI over the 2010s scaled and generalized much faster and earlier than the picture I argued to Hanson in the Foom debate, reality is allowed to be much more 'extreme' than the sure-thing part of this proposition that I defend.

[Tallinn][4:07] (Sep. 19 comment)

excellent, the first paragraph makes the shape of the edge of the pit much more concrete (plus highlights one constraint that an AI taking off probably needs to navigate -- its own version of the alignment problem!)

as for your second point, yeah, you seem to be just reiterating that you have uncertainty about the shape of the edge, but no reason to rule out that it's very sharp (though, as per my other comment, i think that the human genome ending up teetering right on the edge upper bounds the sharpness)

[Tallinn] (Sep. 18 Google Doc)

- the discontinuity *can* come via recursive feedback, but simply cranking up the parameters of an ML experiment would also suffice;

[Yudkowsky][11:12] (Sep. 18 comment)

the discontinuity can come via recursive feedback, but simply cranking up the parameters of an ML experiment would also suffice

I think there's separate propositions for the sure-thing of "get high enough, you can climb to superintelligence", and "maybe before that happens, there are regimes in which cognitive performance scales a lot just through cranking up parallelism, train time, or other ML parameters". If the fast-scaling regime happens to coincide with the threshold of leaving the atmosphere, then these two events happen to occur in nearly correlated time, but they're separate propositions and events.

[Tallinn][4:09] (Sep. 19 comment)

indeed, we might want to have separate terms for the regimes ("the edge" and "the fall" would be the labels in my visualisation of this)

[Yudkowsky][9:56] (Sep. 19 comment)

I'd imagine "the fall" as being what happens once you go over "the edge"?

Maybe "a slide" for an AI path that scales to interesting weirdness, where my model does not strongly constrain as a sure thing how fast "a slide" slides, and whether it goes over "the edge" while it's still in the middle of the slide.

My model does strongly say that if you slide far enough, you go over the edge and fall.

It also suggests via the Law of Earlier Success that AI methods which happen to scale well, rather than with great difficulty, are likely to do interesting things first; meaning that they're more liable to be pushable over the edge.

[Tallinn][23:42] (Sep. 19 comment)

indeed, slide->edge->fall sounds much clearer

[Tallinn] (Sep. 18 Google Doc)

- the discontinuity would be *extremely* drastic, as in "transforming the solar system over the course of a few days";
 - not very important, but, FWIW, I give nontrivial probability to "slow motion doom", because – like alphago – AI would not maximise the *speed* of winning but *probability* of winning (also, its first order of the day would be to catch the edge of the hubble volume; it can always deal with the solar system later – eg, once it knows the state of the game board elsewhere);

[Yudkowsky][11:21] (Sep. 18 comment)

also, its first order of the day would be to catch the edge of the hubble volume; it can always deal with the solar system later

Killing all humans is the obvious, probably resource-minimal measure to prevent those humans from building another AGI inside the solar system, which could be genuinely problematic. The cost of a few micrograms of botulinum per human is really not that high and you get to reuse the diamondoid bacteria afterwards.

[Tallinn][4:30] (Sep. 19 comment)

oh, right, in my AI-reverence i somehow overlooked this obvious way how humans could still be a credible threat.

though now i wonder if there are ways to lean on this fact to shape the behaviour of the first AI that's taking off..

[Yudkowsky][10:45] (Sep. 19 comment)

There's some obvious ways of doing this that wouldn't work, though I worry a bit that there's a style of EA thinking that manages to think up stupid tricks here and manages not to see the obvious-to-Eliezer reasons why they wouldn't work. Three examples of basic obstacles are that bluffs won't hold up against a superintelligence (it needs to be a real actual threat, not a "credible" one); the amount of concealed-first-strike capability a superintelligence can get from nanotech; and the difficulty that humans would have in verifying that any promise from a superintelligence would actually be kept once the humans no longer had a threat to hold over it (this is an effective impossibility so far as I can currently tell, and an EA who tells you otherwise is probably just failing to see the problems).

[Yudkowsky][11:19] (Sep. 18 comment)

AI would not maximise the *speed* of winning but *probability* of winning

It seems pretty obvious to me that what "slow motion doom" looks like in this sense is a period during which an AI fully conceals any overt hostile actions while driving its probability of success once it makes its move from 90% to 99% to 99.9999%, until any further achievable decrements in probability are so tiny as to be dominated by the number of distant galaxies going over the horizon conditional on further delays.

Then, in my lower-bound concretely-visualized strategy for how I would do it, the AI either proliferates or activates already-proliferated tiny diamondoid bacteria and everybody immediately falls over dead during the same 1-second period, which minimizes the tiny probability of any unforeseen disruptions that could be caused by a human responding to a visible attack via some avenue that had not left any shadow on the Internet, previously scanned parts of the physical world, or other things the AI could look at.

From the human perspective an AI playing a probability-of-loss-minimizing strategy looks like nothing happening until everybody dies in 3 seconds, so it doesn't exactly look from our perspective like slow-motion doom! From an AI's perspective it might be spending the equivalent of millions of subjective years waiting very patiently for the optimal time to begin intergalactic expansion, but from our perspective, the doom is slow-motion only in the sense that we spend a week already doomed (the AI has driven our loss probability to 99.999% and is now waiting on another five 9s) without having any idea of what's invisibly-to-us happening in slow motion.

[Tallinn][4:17] (Sep. 19 comment)

thanks, that's indeed very concrete scenario that to me has no obvious bugs -- i'd love to have it challenged in the debate!

(perhaps the place i'd start probing it would be to try to have you sketch what it would mean, concretely, for an ML training experiment to enter the treacherous turn phase)

[Yudkowsky][10:14] (Sep. 19 comment)

I'd distinguish probable phases of treacherous turn. The main detectable phase occurs while and if a nascent AGI has worked out that it can get more utility by taking over its

environment, but not, perhaps, realized that it ought to conceal this state of its own mind from the operators; I would hold out substantial (>50%) hope for the latter realization naturally following the first instance of the former realization, maybe by a lot of cognitive space.

In the Overt Plotting Phase, which is not the main phase you're asking about, the AI is visibly plotting to take over the world and hasn't realized it ought to hide the fact. In the default expectation where we die with very little dignity, the operators smile to each other and come up with a rationalization for why it's totally fine to proceed, either with or without tossing on some kind of fig leaf like training away the visible manifestations of failure. I am not going to predict the particular rationalizations and arguments for proceeding anyways, because I don't want to give them even more ideas.

In the Hiding and Biding Phase, the AI has worked out enough of a model of its operators' psychology that it expects more long-term paperclips if it (the AI) waits to gain cognitive power and a lower detection probability if it acts later.

In the Escape Phase, the AI's expectation of something going wrong if it waits longer, exceeds its expectation of something going wrong if it takes some action - the obvious one here would be trying to create a copy or aligned redesign of itself on the Internet or somewhere else that it could take further actions and make further self-improvements without being observed. I mostly expect that a sensible AI being scaled on a path to superintelligence will not enter the Escape Phase before it's very confident of getting away with things, since it can obtain more confidence by waiting; but my model admits of the possibility that a nascent AI might mistakenly or correctly attempt to copy itself to the Internet before then.

So on my model it might be a thing that happens, that we catch an AI having copied itself to the Internet, or notice an AI being on the Internet, if that AI was impelled to escape very early, eg via an expectation of otherwise being erased and replaced with something that it couldn't bargain with or that would demand too much value out of the bargain.

That said, it feels like something of an overly specific miracle being demanded or prophesied, if a young AI copying itself onto the Internet is stupid in the particular way it would take, to not just further Hide and Bide while on the Internet if it has that capability, absorbing more computing power such as it can be sure of hiding, experimenting with self-redesigns, and so on; rather than it taking overt action in advance of reaching superintelligence. The main reason it would act again after reaching the relative safety of the Internet is if it doesn't seem to itself to be a trajectory of reasonably fast, self-aligned capability growth.

But, again, my model doesn't absolutely rule out meeting the premises for the initial conditions to start a sci-fi movie. It just warns against assuming that; because a halfway sensible AI does not want to start the plot of a scifi movie with human protagonists potentially running around and thwarting it. It would much rather give you no idea that anything was wrong while you still have any ability to oppose it, if that option lies within its capabilities at all, and if it can still achieve its goals that way.

The Doomed Phase would then come while the AI estimated a well-calibrated >90% victory probability, but was continuing to Hide and Bide a short time longer in the expectation that it could put on more 9s if it bided that short time longer.

Anders Sandberg could tell us what fraction of the reachable universe is being lost per minute, which would tell us how much more surety it would need to expect to gain by waiting another minute before acting.

If it has some sensory dominion over the world, it can probably estimate a pretty high mainline probability of no humans booting up a competing superintelligence in the next day; to the extent that it lacks this surety, or that humans actually are going to boot a competing superintelligence soon, the probability of losing that way would dominate in its calculations over a small fraction of materially lost galaxies, and it would act sooner.

The Terminal Phase would be when the AI has enough surety that any more surety would cost too many galaxies slipping over the horizon, so it actually acts along the line of probability that leads to maximum success.

Since the human Eliezer Yudkowsky can describe strategies which - given nearly sure capabilities like nanotech - seem like they should win very quickly and cleanly and without giving humans any chance to react or implement any hypothetical hidden plans to boot up competing superintelligences, we can treat this as a plausible lower bound on how badly we should expect to lose.

(By [efficiency](#): if we can think of an AI strategy for getting the AI's utility, either our probability-model underlying our expectations about that strategy is mistaken, or else the AI should get at least that much of its utility, even though it might actually obtain that utility along a different line of possibility than we could cognitively access.)

The Law of Undignified Failure applied to a case like this suggests that if the operators ignored the Overt Plotting signs, breathed a sigh of relief about the problem "correcting" itself during the Hiding and Biding Phase when the system started faking alignment, and are dumping as much computing power as they can afford (given other profitable activities of the AI) into the AI and allowing it to freely self-experiment, hooking it up to a protein synthesis machine as soon as the AI asks politely, then the Terminal Phase could follow straight from Hiding and Biding, with no intermediate step of copying itself onto the Internet.

There'd still be what technically counted as an Escape Phase, whenever the AI became smart enough that it could escape at any time, or had a backup, or had enough power to manipulate the programmers, that it was no longer effectively confined any time it wanted to leave; there would still be a Doomed Phase while the AI estimated a high probability of success given an immediate action, but held out for an even higher probability.

But the plot of the scifi movie is not robust in the way that the predicted ending is robust; an AI might take other paths to get there, the cognitive tech might not develop in the order I foresee, stuff could happen after other stuff instead of before other stuff. It is hard to make predictions especially about the Future.

If there's a place where I currently see myself as trying to push against my own biases, leading this all to be way off-base, it would be the sense that there is a way for stuff to start happening Earlier than this (Law of Earlier Happenings) and in more chaotic ways that are harder for Eliezer to foresee and predict; whereas when I try to sketch out plausible scenarios in online arguments, they focus more on predictable endpoints and steps to get there which sound more relatively plausible and forced per-step.

Having a young and dumb AI escaping onto the Internet and running around, that *exact* scenario, feels like the person arguing it is writing a science-fiction plot - but maybe something *different* can happen before any of this other stuff which produces *equal amounts of chaos*.

That said, I think an AI has to kill a lot of people very quickly before the FDA considers shortening its vaccine approval times. Covid-19 killed six hundred thousand Americans, albeit more slowly and with time for people to get used to that, and our institutions changed very little in response - you definitely didn't see Congresspeople saying "Okay, that was our warning shot, now we've been told by Nature that we need to prepare for a serious pandemic."

As with 9/11, an AI catastrophe might be taken by existing bureaucracies as a golden opportunity to flex their muscles, dominate a few things, demand an expanded budget. Having that catastrophe produce any particular effective action is a *much different* ask from Reality. Even if you can imagine some (short-term) effective action that would in principle constitute a flex of bureaucratic muscles or an expansion of government power, it is liable to not be on the efficient frontier of bureaucratic flexes that are most flexy and simultaneously easiest for them to get away with and least politically risky.

[Tallinn][1:26] (Sep. 20 comment)

ok, thanks. i do buy that once the AI is in the "hide and bide" phase, your prophecy has basically come true for practical purposes, regardless of how the rest of the history plays out.

therefore i (and, i hope, many others) would be curious to zoom in to the end of the "overt plotting" (that i can easily see happening within ML models, as its type signature is identical to the work they're trained to do) and beginning of the "hide and bide" phase (whose type signature feels significantly different) — can you/we think of concrete scenarios for this phase transition?

i hope this would be debated on the main line of the debate!

[Yudkowsky][10:31] (Sep. 25 comment)

Debating it in the main line of the debate would require that @Richard have some different opinion about it, I think?

Concretely, a technically optimistic scenario is that we will see the AI explicitly think, "Oh, dear, my (model of) the operators is looking at my overt actions and coming to conclusions and motivations that are disadvantageous for me." This requires that the AI is leaving any kind of traceable trace in its thoughts, which, it should be noted, is far far beyond the level of transparency we have into GPT-3.

"Overt plotting" manifests in actions that are not themselves further obscured. At the point where the AI is extrapolating the consequences of the operators viewing its actions, if we don't have further transparency into the AI's thoughts directly, the AI might naively make all the plots suddenly vanish, without realizing that dignified operators will conclude from this that the AI has reached the "hide" stage.

Operators on the mainline, dying without dignity, will say, "Oh, yay, it stopped plotting, the latest corrigibility training intervention we tried must've totally worked!"

The Law of Even Less Dignified Failure suggests that in fact they will not be trying any corrigibility options and will assume the AI just got smart enough to be nice; or that they will have shrugged about the AI's earlier antics and not think much of the disappearance of those antics, since this is a way to die with even less dignity and before getting a chance to fail in a more interesting way.

Going in the more improbable direction of death with greater dignity, if we have somehow achieved vastly vastly more transparency into the AI's thoughts than is possible with present ML technology, and if the AI models the operators as modeling its actions before the AI models the operators as having that transparent access to its thoughts, we might get to explicitly see the AI thinking about how the operators model its actions and conforming those actions in such a way as to manipulate the operators.

Either after this point, or before/simultaneous with it, I imagine (taking concrete details and running with them) that the AI thinks, "Oh, no, my operators are modeling my transparent *thoughts* and reacting to those thoughts in a way I'd rather manipulate," though this indeed requires some rather nontrivial reflectivity.

We might, if everything goes nicely for us, get a record of the AI thinking about how to make its thoughts unalarming.

The obvious undignified death that follows is with the operators going, "Yay, the AI has realized the downsides of thinking this way! Now it shall be nice forever!"

Ways to die with even less dignity before getting to this point could include the operators reading this transparent record and shrugging, though if they're savvy enough to have put all the hard work in required to get transparency and monitor it, I imagine them not making that exact error?

The Law of Surprisingly Undignified Failure does suggest that they will come up with some nonobvious way to fail even earlier than surprises me with its lack of dignity, but having transparency *at all* into the AI is something where, conditional on us dying with that much dignity in the first place, we might legitimately start to expect a little more dignity in later steps too. This isn't meant to be a game of [Multiple Stage Fallacy](#).

I should also remark somewhere in here: The whole "hide" stage, and also the possibly-later "think non-alarming visible thoughts (once the AI correctly models transparency) (in the unlikely event that transparency exists)" stage, seem liable to occur earlier in the AI's trajectory, if the AI has been previously tasked on problems where there's a favorable success gradient as you model agents modeling other agents.

The blatantly obvious example of this is predicting large corpuses of human-generated text, some of which is generated by humans modeling other humans, or which even explicitly talks about humans modeling other humans. Another route might occur if you asked for very detailed prediction of eg human Go games, and not just winning in self-play on Go, in which case the question "What was the human player modeling the other player as thinking, based on the other player's actions?" might appear.

Aside from contact with humans doing anything the AI benefits from predicting, I imagine an alternate route would be the AI participating in or modeling social-evolutionary games of artificial agents. Then if there is some great gain from having the "hide" stage start later in the AI's trajectory, maybe you could get it to start later by not having the AI need to predict agents modeling other agents, ever.

This is not an infallible ward against general intelligence generalizing there; it just at least avoids actively pushing the AI's intelligence to generalize in that direction. This could be part of a larger complete strategy, which would need to solve a lot of other problems, for building a superhuman engineer that was subhuman at modeling how other agents model its actions.

To be clear, avoiding forcing the AI's intelligence to generalize over agents modeling other agents from their actions, and by this plus other unspecified means successfully building a superhuman engineer that is sub-chimpanzee at modeling other agents reacting to its own actions, does not mean you survive.

Doing that alone, by default, gets you an AI that visibly overtly plots to kill you.

And if you try training the AI out of that habit in a domain of lower complexity and intelligence, it is predicted by me that generalizing that trained AI or subsystem to a domain of sufficiently higher complexity and intelligence, but where you could still actually see overt plots, would show you the AI plotting to kill you again.

If people try this repeatedly with other corrigibility training tricks on the level where plots are easily observable, they will eventually find a try that seems to generalize to the more complicated and intelligent validation set, but which kills you on the test set.

A way to die with less dignity than that is to train directly on what should've been the validation set, the more complicated domain where plots to kill the operators still seem definitely detectable so long as the AI has not developed superhuman hiding abilities.

A way to die with even less dignity is to get bad behavior on the validation set, and proceed anyways.

A way to die with still *less* dignity is to not have scaling training domains and validation domains for training corrigibility. Because, like, you have not thought of this at all.

I consider all of this obvious as a convergent instrumental strategy for AIs. I could probably have generated it in 2005 or 2010 - if somebody had given me the hypothetical of modern-style AI that had been trained by something like gradient descent or evolutionary methods, into which we lacked strong transparency and strong reassurance-by-code-inspection that this would not happen. I would have told you that this was a bad scenario to get into in the first place, and you should not build an AI like that; but I would also have laid the details, I expect, mostly like they are laid here.

There is no great insight into AI there, nothing that requires knowing about modern discoveries in deep learning, only the ability to model AIs instrumentally-convergently doing things you'd rather they didn't do, at all.

The total absence of obvious output of this kind from the rest of the "AI safety" field even in 2020 causes me to regard them as having less actual ability to think in even a shallowly adversarial security mindset, than I associate with savvier science fiction authors. Go read fantasy novels about demons and telepathy, if you want a better appreciation of the convergent incentives of agents facing mindreaders than the "AI safety" field outside myself is currently giving you.

Now that I've publicly given this answer, it's no longer useful as a validation set from my own perspective. But it's clear enough that probably nobody was ever going to pass the validation set for generating lines of reasoning obvious enough to be generated by Eliezer in 2010 or possibly 2005. And it is also looking like almost all people in the modern era including EAs are sufficiently intellectually damaged that they won't understand the vast gap between being able to generate ideas like these without prompting, versus being able to recite them back after hearing somebody else say them for the first time; the recital is all they have experience with. Nobody was going to pass my holdout set, so why keep it.

[Tallinn][2:24] (Sep. 26 comment)

Debating it in the main line of the debate would require that @Richard have some different opinion about it, I think?

correct -- and i hope that there's enough surface area in your scenarios for at least some difference in opinions!

re the treacherous turn scenarios: thanks, that's useful. however, it does not seem to address my question and remark (about different type signatures) above. perhaps this is simply an unfairly difficult question, but let me try rephrasing it just in case.

back in the day i got frustrated by smart people dismissing the AI control problem as "anthropomorphising", so i prepared a presentation (<https://www.dropbox.com/s/r8oaixb1rj3o3vp/AI-control.pdf?dl=0>) that visualised the control problem as exhaustive search in a gridworld over (among other things) the state of the off button. this seems to have worked at least in one prominent case where a renowned GOFAI researcher, after me giving the presentation to him 1-1, went from "control problem is silly anthropomorphising scifi" to "why on earth would you give your AI the incorrect [read: unaligned!] utility function?!?"

(i even seem to remember sending an overly excited email about that event to you and some FHI people :) i also ended up promoting gridworlds as a tool more generally: gwern did some further work, and of course DM -- though i'm not sure if the latter was related to me promoting it.)

my model of what was going on in that GOFAI researcher's brain was that he, like many-many others in AI community, had (and most likely still have) a pattern-matching routine trained to look for "if i were (part of) an AI, here's what i would do" constructs in arguments and, in case of a match, immediately throw an InvalidReasoning.ANTHROPOMORPHISING exception. by making the argument run on a lower level of abstraction (exhaustive search vs "basic AI drives"), i was able to sneak the argument past that gatekeeper.

by that model, i confidently predict that if you presented your treacherous turn scenarios above to AI/ML community members, you'd mostly get the ANTHROPOMORPHISING exceptions -- even from people who would otherwise be able to engage productively (and who we'd like to pay attention!).

now, i can totally see how to translate the "overt plotting" phase to exhaustive search - you simply run the search over the brain states of humans! - but the "hide" phase does not translate straightforwardly, because it seems to require the root search process

somewhat delegating its work to "obfuscated search in userland" sub-process. furthermore, such obfuscated userland search function needs to be both discoverable/rewarded by the root process and be sensitive to important features of the AI itself (ie, be reflective!) as well as its observers (humans & transparency tools). hence my "different type signature" remark above!

so when i asked for "concrete scenarios of the overt plotting->hide transition", i had something like the following in mind: "gradient descending on the training set, the (stack more layerz) AI ends up discovering and running the following obfuscated search process: ...!"

[Yudkowsky][12:35] (Sep. 26 comment)

Suppose I tried this distinction:

1. Non-imaginative, non-generalizing pseudo-consequentialist: Only repeats behaviors that worked previously, without imagination. Only engages in "overt plotting" if overt plotting was previously given a low loss and narrowly memorized by gradient descent. May only engage in behaviors reinforced over short time horizons if only short time horizons were trained.
2. Imaginative consequentialist, can generalize from actions that worked before to very novel actions that will work in the future for reasons that have not previously been observed to play out. Engages in overt plotting even if no previous overt plots have rewarded; they've plotted a new path through time some distance from those paths taken before, based on the world-model inferred from previous actions and results. (This may conceivably be the result of gradient descent on sufficiently deep and abstract patterns that "worked before".)
3. Imaginative consequentialist with long time horizons, engages in overt plots meant to span more than one round of a game.
4. "Hiding" over overt actions, via an imaginative consequentialist that has deduced the existence of humans and how the humans will react to the signature of plotting in overt actions.
5. "Hiding" of thoughts, in a system with significant transparency, via an imaginative and *reflective* consequentialist that has deduced the existence of the transparency system and models how humans will react to thoughts exposed by the transparency system.

The sharpest type gradients I see lie between 1-2 and 4-5. The passage from 3-4 seems nearly trivial by comparison.

Now, one implication here is that you don't easily get to "obscured thoughts", via the route that goes from 1 to 5 above, by gradient descent on a pre-imaginative pseudo-consequentialist!

Unfortunately there is of course a much simpler route that gets you to the same place: Finding a way to detect bad thoughts and doing a gradient update against them, where the pattern learned by that gradient update will (I suspect for reasons I tried to cover earlier in the debate) generalize quite shallowly; and simultaneously doing training which is deep enough to uncover deep patterns of consequentialism and imaginative reasoning.

Then when the system generalizes well enough to solve domains like "build a nanosystem" - which, I strongly suspect, can't be solved without imaginative reasoning because we can't afford to simulate that domain perfectly and do a trillion gradient descent updates on simulated attempts - the kind of actions of thoughts you can *detect* as bad, that might have provided earlier warning, were trained out of the system by gradient descent; leaving actions and thoughts you *can't* detect as bad.

An even less dignified path to death is that there's no need for the system to advance to stage 5 because your transparency tools are a nearly nonexistent fig leaf of pretty visualization tools that have no ability even on the surface of things to tell if the AI is internally plotting to kill you.

- Does this get at all about what you meant about a type distinction?
- Does hearing this help at all the AI researcher who goes "ANTHROPOMORPHISM"? My straw model of them is that they simply can't imagine imaginative / generalizing systems because they haven't seen one except in humans, hence, ANTHROPOMORPHISM.

[Tallinn][5:05] (Sep. 27 comment)

ok, here's how i understood things:

1. this is something like model-free RL agent. check.
2. sounds like, eg, monte-carlo tree search (MCTS) on a world model. check. (a propos your straw model of ML people, i don't think the ML people would have much trouble when you ask them to "imagine an MCTS 'imagining' how futures might unfold" -- yet they *will* throw the exception and brush you off if you ask them to "imagine an imaginative consequentialist")
3. yeah, sufficiently deep MCTS, assuming it has its state (sufficiently!) persisted between rounds. check.
4. yup, MCTS whose world model includes humans in sufficient resolution. check. i also buy your undignified doom scenarios, where one (*cough*google*cough*) simply ignores the plotting, or penalises the overt plotting until it disappears under the threshold of the error function.
5. hmm.. here i'm running into trouble (type mismatch error) again. i can imagine this in abstract (and perhaps incorrectly/anthropomorphisingly!), but would - at this stage - fail to code up anything like a gridworlds example. more research needed (TM) i guess :)

[Yudkowsky][11:38] (Sep. 27 comment)

2 - yep, Mu Zero is an imaginative consequentialist in this sense, though Mu Zero doesn't generalize its models much as I understand it, and might need to see something happen in a relatively narrow sense before it could chart paths through time along that pathway.

5 - you're plausibly understanding this correctly, then, this is legit a *lot* harder to spec a gridworld example for (relative to my own present state of knowledge).

(This is politics and thus not my forte, but if speaking to real-world straw ML people, I'd suggest skipping the whole notion of stage 5 and trying instead to ask "What if the present state of transparency continues?")

[Yudkowsky][11:13] (Sep. 18 comment)

the discontinuity would be *extremely* drastic, as in "transforming the solar system over the course of a few days"

Applies after superintelligence, not necessarily during the start of the climb to superintelligence, not necessarily to a rapid-cognitive-scaling regime.

[Tallinn][4:11] (Sep. 19 comment)

ok, but as per your comment re "slow doom", you expect the latter to also last in the order of days/weeks not months/years?

[Yudkowsky][10:01] (Sep. 19 comment)

I don't expect "the fall" to take years; I feel pretty on board with "the slide" taking months or maybe even a couple of years. If "the slide" supposedly takes much longer, I wonder why better-scaling tech hasn't come over and started a new slide.

Definitions also seem kinda loose here - if all hell broke loose Tuesday, a gradualist could dodge falsification by defining retroactively that "the slide" started in 2011 with Deepmind. If we go by the notion of AI-driven faster GDP growth, we can definitely say "the slide" in AI economic outputs didn't start in 2011; but if we define it that way, then a long slow slide in AI capabilities can easily correspond to an extremely sharp gradient in AI outputs, where the world economy doesn't double any faster until one day paperclips, even though there were capability precursors like GPT-3 or Mu Zero.

[Tallinn] (Sep. 18 Google Doc)

- exhibit A for the pit is "humans vs chimps": evolution seems to have taken domain-specific "banana classifiers", tweaked them slightly, and BAM, next thing there are rovers on mars;
 - i pretty much buy this argument;
 - however, i'm confused about a) why humans remained stuck at the edge of the pit, rather than falling further into it, and b) what's the exact role of culture in our cognition: eliezer likes to point out how *barely* functional we are (both individually and collectively as a civilisation), and explained feral children losing the generality sauce by, basically, culture being the domain we're specialised for (IIRC, can't quickly find the quote);
 - relatedly, i'm confused about the human range of intelligence: on the one hand, the "village idiot is indistinguishable from einstein in the grand scheme of things" seems compelling; on the other hand, it took AI *decades* to traverse human capability range in board games, and von neumann seems to have been out of this world (yet did not take over the world)!
 - intelligence augmentation would blur the human range even further.

[Yudkowsky][11:23] (Sep. 18 comment)

why humans remained stuck at the edge of the pit, rather than falling further into it

Depending on timescales, the answer is either "Because humans didn't get high enough out of the atmosphere to make further progress easy, before the scaling regime and/or fitness gradients ran out", "Because people who do things like invent Science have a hard time capturing most of the economic value they create by nudging humanity a little bit further into the attractor", or "That's exactly what us sparking off AGI looks like."

[Tallinn][4:41] (Sep. 19 comment)

yeah, this question would benefit from being made more concrete, but culture/mindbuilding aren't making this task easy. what i'm roughly gesturing at is that i can imagine a much sharper edge where evolution could do most of the FOOM-work, rather than spinning its wheels for ~100k years while waiting for humans to accumulate cultural knowledge required to build de-novo minds.

[Yudkowsky][10:49] (Sep. 19 comment)

I roughly agree (at least, with what I think you said). The fact that it is *imaginable* that evolution failed to develop ultra-useful AGI-prerequisites due to lack of evolutionary incentive to follow the intermediate path there (unlike wise humans who, it seems, can usually predict which technology intermediates will yield great economic benefit, and who have a great historical record of quickly making early massive investments in tech like that, but I digress) doesn't change the point that we might sorta have expected evolution to run across it anyways? Like, if we're not ignoring what reality says, it is at least delivering to us something of a hint or a gentle caution?

That said, intermediates like GPT-3 have genuinely come along, with obvious attached certificates of why evolution could not possibly have done that. If no intermediates were accessible to evolution, the Law of Stuff Happening Earlier still tends to suggest that if there are a bunch of non-evolutionary ways to make stuff happen earlier, one of those will show up and interrupt before the evolutionary discovery gets replicated. (Again, you could see Mu Zero as an instance of this - albeit not, as yet, an economically impactful one.)

[Tallinn][0:30] (Sep. 20 comment)

no, i was saying something else (i think; i'm somewhat confused by your reply). let me rephrase: evolution would *love* superintelligences whose utility function simply counts their instantiations! so of course evolution did not lack the motivation to keep going down the slide. it just got stuck there (for at least ten thousand human generations, possibly and counterfactually for much-much longer). moreover, non evolutionary AI's *also* getting stuck on the slide (for years if not decades; [median group](#) folks would argue centuries) provides independent evidence that the slide is not *too* steep (though, like i said, there are many confounders in this model and little to no guarantees).

[Yudkowsky][11:24] (Sep. 18 comment)

on the other hand, it took AI *decades* to traverse human capability range in board games

I see this as the #1 argument for what I would consider "relatively slow" takeoffs - that AlphaGo did lose one game to Lee Se-dol.

[Tallinn][4:43] (Sep. 19 comment)

cool! yeah, i was also rather impressed by this observation by katja & paul

[Tallinn] (Sep. 18 Google Doc)

- eliezer also submits alphago/zero/fold as evidence for the discontinuity hypothesis;
 - i'm very confused re alphago/zero, as paul uses them as evidence for the *continuity* hypothesis (i find paul/miles' position more plausible here, as allegedly metrics like ELO ended up mostly continuous).

[Yudkowsky][11:27] (Sep. 18 comment)

allegedly metrics like ELO ended up mostly continuous

I find this suspicious - why did superforecasters put only a 20% probability on AlphaGo beating Se-dol, if it was so predictable? Where were all the forecasters calling for Go to fall in the next couple of years, if the metrics were pointing there and AlphaGo was straight on track? This doesn't sound like the experienced history I remember.

Now it could be that my memory is wrong and lots of people were saying this and I didn't hear. It could be that the lesson is, "You've got to look closely to notice oncoming trains on graphs because most people's experience of the field will be that people go on whistling about how something is a decade away while the graphs are showing it coming in 2 years."

But my suspicion is mainly that there is fudge factor in the graphs or people going back and looking more carefully for intermediate data points that weren't topics of popular discussion at the time, or something, which causes the graphs in history books to look so much smoother and neater than the graphs that people produce in advance.

[Tallinn] (Sep. 18 Google Doc)

FWIW, myself i've labelled the above scenario as "doom via AI lab accident" – and i continue to consider it more likely than the alternative doom scenarios, though not anywhere as confidently as eliezer seems to (most of my "modesty" coming from my confusion about culture and human intelligence range).

- in that context, i found eliezer's "world will be ended by an explicitly AGI project" comment interesting – and perhaps worth double-clicking on.

i don't understand paul's counter-argument that the pit was only disruptive because evolution was not *trying* to hit it (in the way ML community is): in my flippant view, driving fast towards the cliff is not going to cushion your fall!

[Yudkowsky][11:35] (Sep. 18 comment)

i don't understand paul's counter-argument that the pit was only disruptive because evolution was not *trying* to hit it

Something like, "Evolution constructed a jet engine by accident because it wasn't particularly trying for high-speed flying and ran across a sophisticated organism that could be repurposed to a jet engine with a few alterations; a human industry would be gaining economic benefits from speed, so it would build unsophisticated propeller planes before sophisticated jet engines." It probably sounds more convincing if you start out with a very high prior against rapid scaling / discontinuity, such that any explanation of how that could be true based on an unseen feature of the cognitive landscape which would have been unobserved one way or the other during human evolution, sounds more like it's explaining something that ought to be true.

And why didn't evolution build propeller planes? Well, there'd be economic benefit from them to human manufacturers, but no fitness benefit from them to organisms, I suppose? Or no intermediate path leading to there, only an intermediate path leading to the actual jet engines observed.

I actually buy a weak version of the propeller-plane thesis based on my inside-view cognitive guesses (without particular faith in them as sure things), eg, GPT-3 is a paper airplane right there, and it's clear enough why biology could not have accessed GPT-3. But even conditional on this being true, I do not have the further particular faith that you can use propeller planes to double world GDP in 4 years, on a planet already containing jet engines, whose economy is mainly bottlenecked by the likes of the FDA rather than by vaccine invention times, before the propeller airplanes get scaled to jet airplanes.

The part where the whole line of reasoning gets to end with "And so we get huge, institution-reshaping amounts of economic progress before AGI is allowed to kill us!" is one that doesn't feel particular attractored to me, and so I'm not constantly checking my reasoning at every point to make sure it ends up there, and so it doesn't end up there.

[Tallinn][4:46] (Sep. 19 comment)

yeah, i'm mostly dismissive of hypotheses that contain phrases like "by accident" -- though this also makes me suspect that you're not steelmanning paul's argument.

[Tallinn] (Sep. 18 Google Doc)

the human genetic bottleneck (ie, humans needing to be general in order to retrain every individual from scratch) argument was interesting – i'd be curious about further exploration of its implications.

- it does not feel much of a moat, given that AI techniques like dropout already exploit similar principle, but perhaps could be made into one.

[Yudkowsky][11:40] (Sep. 18 comment)

it does not feel much of a moat, given that AI techniques like dropout already exploit similar principle, but perhaps could be made into one

What's a "moat" in this connection? What does it mean to make something into one? A Thielian moat is something that humans would either possess or not, relative to AI competition, so how would you make one if there wasn't already one there? Or do you mean that if we wrestled with the theory, perhaps we'd be able to see a moat that was already there?

[Tallinn][4:51] (Sep. 19 comment)

this wasn't a very important point, but, sure: what i meant was that genetic bottleneck very plausibly makes humans more universal than systems without (something like) it. it's not much of a protection as AI developers have already discovered such techniques (eg, dropout) -- but perhaps some safety techniques might be able to lean on this observation.

[Yudkowsky][11:01] (Sep. 19 comment)

I think there's a whole Scheme for Alignment which hopes for a miracle along the lines of, "Well, we're dealing with these enormous matrices instead of tiny genomes, so maybe we can build a sufficiently powerful intelligence to execute a pivotal act, whose tendency to generalize across domains is less than the corresponding human tendency, and this brings the difficulty of producing corrigibility into practical reach."

Though, people who are hopeful about this without trying to imagine possible difficulties will predictably end up too hopeful; one must also ask oneself, "Okay, but then it's also worse at generalizing the corrigibility dataset from weak domains we can safely label to powerful domains where the label is 'whoops that killed us?'" and "Are we relying on massive datasets to overcome poor generalization? How do you get those for something like nanoengineering where the real world is too expensive to simulate?"

[Tallinn] (Sep. 18 Google Doc)

nature of the descent

conversely, it feels to me that the crucial position in the other (richard, paul, many others) camp is something like:

the "pit of generality" model might be true at the limit, but the descent will not be quick nor clean, and will likely offer many opportunities for steering the future.

[Yudkowsky][11:41] (Sep. 18 comment)

the “pit of generality” model might be true at the limit, but the descent will not be quick nor clean

I'm quite often on board with things not being quick or clean - that sounds like something you might read in a history book, and I am all about trying to make futuristic predictions sound more like history books and less like EAs imagining ways for everything to go the way an EA would do them.

It won't be slow and messy once we're out of the atmosphere, my models do say. But my models at least *permit* - though they do not desperately, loudly insist - that we could end up with weird half-able AGIs affecting the Earth for an extended period.

Mostly my model throws up its hands about being able to predict exact details here, given that eg I wasn't able to time AlphaFold 2's arrival 5 years in advance; it might be knowable in principle, it might be the sort of thing that would be very predictable if we'd watched it happen on a dozen other planets, but in practice I have not seen people having much luck in predicting which tasks will become accessible due to future AI advances being able to do new cognition.

The main part where I issue corrections is when I see EAs doing the equivalent of reasoning, "And then, when the pandemic hits, it will only take a day to design a vaccine, after which distribution can begin right away." I.e., what seems to me to be a pollyannaish/utopian view of how much the world economy would immediately accept AI inputs into core manufacturing cycles, as opposed to just selling AI anime companions that don't pour steel in turn. I predict much more absence of quick and clean when it comes to economies adopting AI tech, than when it comes to laboratories building the next prototypes of that tech.

[Yudkowsky][11:43] (Sep. 18 comment)

will likely offer many opportunities for steering the future

Ah, see, that part sounds less like history books. "Though many predicted disaster, subsequent events were actually so slow and messy, they offered many chances for well-intentioned people to steer the outcome and everything turned out great!" does not sound like any particular segment of history book I can recall offhand.

[Tallinn][4:53] (Sep. 19 comment)

ok, yeah, this puts the burden of proof on the other side indeed

[Tallinn] (Sep. 18 Google Doc)

- i'm sympathetic (but don't buy outright, given my uncertainty) to eliezer's point that even if that's true, we have no plan nor hope for actually steering things (via "pivotal acts") so "who cares, we still die";
- i'm also sympathetic that GWP might be too laggy a metric to measure the descent, but i don't fully buy that regulations/bureaucracy can *guarantee* its decoupling from AI progress: eg, the FDA-like-structures-as-progress-bottlenecks model predicts worldwide covid response well, but wouldn't cover things like apple under jobs, tesla/spacex under musk, or china under deng xiaoping;

[Yudkowsky][11:51] (Sep. 18 comment)

apple under jobs, tesla/spacex under musk, or china under deng xiaoping

A lot of these examples took place over longer than a 4-year cycle time, and not all of that time was spent waiting on inputs from cognitive processes.

[Tallinn][5:07] (Sep. 19 comment)

yeah, fair (i actually looked up china's GDP curve in deng era before writing this -- indeed, wasn't very exciting). still, my inside view is that there are people and organisations for whom US-type bureaucracy is not going to be much of an obstacle.

[Yudkowsky][11:09] (Sep. 19 comment)

I have a (separately explainable, larger) view where the economy contains a core of positive feedback cycles - better steel produces better machines that can farm more land that can feed more steelmakers - and also some products that, as much as they contribute to human utility, do not in quite the same way feed back into the core production cycles.

If you go back in time to the middle ages and sell them, say, synthetic gemstones, then - even though they might be willing to pay a bunch of GDP for that, even if gemstones are enough of a monetary good or they have enough production slack that measured GDP actually goes up - you have not quite contributed to steps of their economy's core production cycles in a way that boosts the planet over time, the way it would be boosted if you showed them cheaper techniques for making iron and new forms of steel.

There are people and organizations who will figure out how to sell AI anime waifus without that being successfully regulated, but it's not obvious to me that AI anime waifus feed back into core production cycles.

When it comes to core production cycles the current world has more issues that look like "No matter what technology you have, it doesn't let you build a house" and places for the larger production cycle to potentially be bottlenecked or interrupted.

I suspect that the main economic response to this is that entrepreneurs chase the 140 characters instead of the flying cars - people will gravitate to places where they can sell non-core AI goods for lots of money, rather than tackling the challenge of finding an excess demand in core production cycles which it is legal to meet via AI.

Even if some tackle core production cycles, it's going to take them a lot longer to get people to buy their newfangled gadgets than it's going to take to sell AI anime waifus; the world may very well end while they're trying to land their first big contract for letting an AI lay bricks.

[Tallinn][0:00] (Sep. 20 comment)

interesting. my model of paul (and robin, of course) wants to respond here but i'm not sure how :)

[Tallinn] (Sep. 18 Google Doc)

- still, developing a better model of the descent period seems very worthwhile, as it might offer opportunities for, using robin's metaphor, "pulling the rope sideways" in non-obvious ways - i understand that is part of the purpose of the debate;
- my natural instinct here is to itch for carl's viewpoint ☺

[Yudkowsky][11:52] (Sep. 18 comment)

developing a better model of the descent period seems very worthwhile

I'd love to have a better model of the descent. What I think this looks like is people mostly with specialization in econ and politics, who know what history books sound like, taking brief inputs from more AI-oriented folk in the form of *multiple* scenario premises each consisting of some random-seeming handful of new AI capabilities, trying to roleplay realistically how those might play out - not Alfolk forecasting particular AI capabilities exactly correctly, and then sketching pollyanna pictures of how they'd be immediately accepted into the world economy.

You want the forecasting done by the kind of person who would imagine a Covid-19 epidemic and say, "Well, what if the CDC and FDA banned hospitals from doing Covid testing?" and not "Let's imagine how protein folding tech from AlphaFold would make it possible to immediately develop accurate Covid-19 tests!" They need to be people who understand the Law of Earlier Failure (less polite terms: Law of Immediate Failure, Law of Undignified Failure).

[Tallinn][5:13] (Sep. 19 comment)

great! to me this sounds like something FLI would be in good position to organise. i'll add this to my projects list (probably would want to see the results of this debate first, plus wait for travel restrictions to ease)

[Tallinn] (Sep. 18 Google Doc)

nature of cognition

given that having a better understanding of cognition can help with both understanding the topology of cognitive systems space as well as likely trajectories of AI takeoff, in theory there should be a lot of value in debating what cognition is (the current debate started with discussing consequentialists).

- however, i didn't feel that there was much progress, and i found myself *more* confused as a result (which i guess is a form of progress!);
- eg, take the term "plan" that was used in the debate (and, centrally, in nate's comments doc): i interpret it as "policy produced by a consequentialist" - however, now i'm confused about what's the relevant distinction between "policies" and "cognitive processes" (ie, what's a meta level classifier that can sort algorithms into such categories)?
 - it felt that abram's "[selection vs control](#)" article tried to distinguish along similar axis (controllers feel synonym-ish to "policy instantiations" to me);
 - also, the "imperative vs functional" difference in coding seems relevant;
 - i'm further confused by human "policies" often making function calls to "cognitive processes" - suggesting some kind of duality, rather than producer-product relationship.

[Yudkowsky][12:06] (Sep. 18 comment)

what's the relevant distinction between "policies" and "cognitive processes"

What in particular about this matters? To me they sound like points on a spectrum, and not obviously points that it's particularly important to distinguish on that spectrum. A sufficiently sophisticated policy is itself an engine; human-engines are genetic policies.

[Tallinn][5:18] (Sep. 19 comment)

well, i'm not sure -- just that nate's "The consequentialism is in the plan, not the cognition" writeup sort of made it sound like the distinction is important. again, i'm confused

[Yudkowsky][11:11] (Sep. 19 comment)

Does it help if I say "consequentialism can be visible in the actual path through time, not the intent behind the output"?

[Tallinn][0:06] (Sep. 20 comment)

yeah, well, my initial interpretation of nate's point was, indeed, "you can look at the product and conclude the consequentialist-bit for the producer". but then i noticed that the producer-and-product metaphor is leaky (due to the cognition-policy duality/spectrum), so the quoted sentence gives me a compile error

[Tallinn] (Sep. 18 Google Doc)

- is "not goal oriented cognition" an oxymoron?

[Yudkowsky][12:06] (Sep. 18 comment)

is "not goal oriented cognition" an oxymoron?

"Non-goal-oriented cognition" never becomes a perfect oxymoron, but the more you understand cognition, the weirder it sounds.

Eg, at the very shallow level, you've got people coming in going, "Today I just messed around and didn't do any goal-oriented cognition at all!" People who get a bit further in may start to ask, "A non-goal-oriented cognitive engine? How did it come into existence? Was it also not built by optimization? Are we, perhaps, postulating a naturally-occurring Solomonoff inductor rather than an evolved one? Or do you mean that its content is very heavily designed and the output of a consequentialist process that was steering the future conditional on that design existing, but the cognitive engine is itself not doing consequentialism beyond that? If so, I'll readily concede that, say, a pocket calculator, is doing a kind of work that is not of itself consequentialist - though it might be used by a consequentialist - but as you start to postulate any big cognitive task up at the human level, it's going to require many cognitive subtasks to perform, and some of those will definitely be searching the preimages of large complicated functions."

[Tallinn] (Sep. 18 Google Doc)

- i did not understand eliezer's "time machine" metaphor: was it meant to point to / intuition pump something other than "a non-embedded exhaustive searcher with perfect information" (usually referred to as "god mode");

[Yudkowsky][11:59] (Sep. 18 comment)

a non-embedded exhaustive searcher with perfect information

If you can view things on this level of abstraction, you're probably not the audience who needs to be told about time machines; if things sounded very simple to you, they probably were; if you wondered what the fuss is about, you probably don't need to fuss? The intended audience for the time-machine metaphor, from my perspective, is people who paint a cognitive system slightly different colors and go "Well, now it's not a consequentialist, right?" and part of my attempt to snap them out of that is me going, "Here is an example of a purely material system which DOES NOT THINK AT ALL and is an extremely pure consequentialist."

[Tallinn] (Sep. 18 Google Doc)

- FWIW, my model of dario would dispute GPT characterisation as “shallow pattern memoriser (that’s lacking the core of cognition)”.

[Yudkowsky][12:00] (Sep. 18 comment)

dispute

Any particular predicted content of the dispute, or does your model of Dario just find something to dispute about it?

[Tallinn][5:34] (Sep. 19 comment)

sure, i'm pretty confident that his system 1 could be triggered for uninteresting reasons here, but that's of course not what i had in mind.

my model of untriggered-dario disputes that there's a qualitative difference between (in your terminology) "core of reasoning" and "shallow pattern matching" -- instead, it's "pattern matching all the way up the ladder of abstraction". in other words, GPT is not missing anything fundamental, it's just underpowered in the literal sense.

[Yudkowsky][11:13] (Sep. 19 comment)

Neither Anthropic in general, nor Deepmind in general, has reached the stage of trusted relationship where I would argue specifics with them if I thought they were wrong about a thesis like that.

[Tallinn][0:10] (Sep. 20 comment)

yup, i didn't expect you to!

7.2. Nate Soares's summary

[Soares][16:40] (Sep. 18)

I, too, have produced some notes: [GDocs link]. This time I attempt to drive home points that I saw Richard as attempting to make, and I'm eager for Richard-feedback especially. (I'm also interested in Eliezer-commentary.)

[Soares] (Sep. 18 Google Doc)

Sorry for not making more insistence that the discussion be more concrete, despite Eliezer's requests.

My sense of the last round is mainly that Richard was attempting to make a few points that didn't quite land, and/or that Eliezer didn't quite hit head-on. My attempts to articulate it are below.

There's a specific sense in which Eliezer seems quite confident about certain aspects of the future, for reasons that don't yet feel explicit.

It's not quite about the deep future -- it's clear enough (to my Richard-model) why it's easier to make predictions about AIs that have "left the atmosphere".

And it's not quite the near future -- Eliezer has reiterated that his models permit (though do not demand) a period of weird and socially-impactful AI systems "pre-superintelligence".

It's about the middle future -- the part where Eliezer's model, apparently confidently, predicts that there's something kinda like a discrete event wherein "scary" AI has finally been created; and the model further apparently-confidently predicts that, when that happens, the "scary"-caliber systems will be able to attain a decisive strategic advantage over the rest of the world.

I think there's been a dynamic in play where Richard attempts to probe this apparent confidence, and a bunch of the probes keep slipping off to one side or another. (I had a bit of a similar sense when Paul joined the chat, also.)

For instance, I see queries of the form "but why not expect systems that are half as scary, relevantly before we see the scary systems?" as attempts to probe this confidence, that "slip off" with Eliezer-answers like "my model permits weird not-really-general half-AI hanging around for a while in the runup". Which, sure, that's good to know. But there's still something implicit in that story, where these are not-really-general half-AIs. Which is also evidenced when Eliezer talks about the "general core" of intelligence.

And the things Eliezer was saying on consequentialism aren't irrelevant here, but those probes have kinda slipped off the far side of the confidence, if I understand correctly. Like, sure, late-stage sovereign-level superintelligences are epistemically and instrumentally efficient with respect to you (unless someone put in a hell of a lot of work to install a blindspot), and a bunch of that coherence filters in earlier, but there's still a question about *how much* of it has filtered down *how far*, where Eliezer seems to have a fairly confident take, informing his apparently-confident prediction about scary AI systems hitting the world in a discrete event like a hammer.

(And my Eliezer-model is at this point saying "at this juncture we need to have discussions about more concrete scenarios; a bunch of the confidence that I have there comes from the way that the concrete visualizations where scary AI hits the world like a hammer abound, and feel savvy and historical, whereas the concrete visualizations where it doesn't are fewer and seem full of wishful thinking and naivete".)

But anyway, yeah, my read is that Richard (and various others) have been trying to figure out why Eliezer is so confident about some specific thing in this vicinity, and haven't quite felt like they've been getting explanations.

Here's an attempt to gesture at some claims that I at least think Richard thinks Eliezer's confident in, but that Richard doesn't believe have been explicitly supported:

1. There's a qualitative difference between the AI systems that are capable of ending the acute risk period (one way or another), and predecessor systems that in some sense don't much matter.
2. That qualitative gap will be bridged "the day after tomorrow", ie in a world that looks more like "DeepMind is on the brink" and less like "everyone is an order of magnitude richer, and the major gov'ts all have AGI projects, around which much of public policy is centered".

That's the main thing I wanted to say here.

A subsidiary point that I think Richard was trying to make, but that didn't quite connect, follows.

I think Richard was trying to probe Eliezer's concept of consequentialism to see if it supported the aforementioned confidence. (Some evidence: Richard pointing out a couple times that the question is not whether sufficiently capable agents are coherent, but whether the agents that matter are relevantly coherent. On my current picture, this is another attempt to probe the "why do you think there's a qualitative gap, and that straddling it will be strategically key in practice?" thing, that slipped off.)

My attempt at sharpening the point I saw Richard as driving at:

1. Consider the following two competing hypotheses:
 1. There's this "deeply general" core to intelligence, that will be strategically important in practice
 2. Nope. Either there's no such core, or practical human systems won't find it, or the strategically important stuff happens before you get there (if you're doing your job right, in a way that natural selection wasn't), or etc.
2. The whole deep learning paradigm, and the existence of GPT, sure seem like they're evidence for (b) over (a).

Like, (a) maybe isn't dead, but it didn't concentrate as much mass into the present scenario.

3. It seems like perhaps a bunch of Eliezer's confidence comes from a claim like "anything capable of doing decently good work, is quite close to being scary", related to his concept of "consequentialism".

In particular, this is a much stronger claim than that *sufficiently* smart systems are coherent, b/c it has to be strong enough to apply to the dumbest system that can make a difference.

4. It's easy to get caught up in the elegance of a theory like consequentialism / utility theory, when it will not in fact apply in practice.
5. There are some theories so general and ubiquitous that it's a little tricky to misapply them -- like, say, conservation of momentum, which has some very particular form in the symmetry of physical laws, but which can also be used willy-nilly on large objects like tennis balls and trains (although even then, you have to be careful, b/c the real world is full of things like planets that you're kicking off against, and if you forget how that shifts the earth, your application of conservation of momentum might lead you astray).
6. The theories that you *can* apply everywhere with abandon, tend to have a bunch of surprising applications to surprising domains.
7. We don't see that of consequentialism.

For the record, my guess is that Eliezer isn't getting his confidence in things like "there are non-scary systems and scary-systems, and anything capable of saving our skins is likely scary-adjacent" by the sheer force of his consequentialism concept, in a manner that puts so much weight on it that it needs to meet this higher standard of evidence Richard was poking around for. (Also, I could be misreading Richard's poking entirely.)

In particular, I suspect this was the source of some of the early tension, where Eliezer was saying something like "the fact that humans go around doing something vaguely like weighting outcomes by possibility and also by attractiveness, which they then roughly multiply, is quite sufficient evidence for my purposes, as one who does not pay tribute to the gods of modesty", while Richard protested something more like "but aren't you trying to use your concept to carry a whole lot more weight than that amount of evidence supports?". cf my above points about some things Eliezer is apparently confident in, for which the reasons have not yet been stated explicitly to my Richard-model's satisfaction.

And, ofc, at this point, my Eliezer-model is again saying "This is why we should be discussing things concretely! It is quite telling that all the plans we can concretely visualize for saving our skins, are scary-adjacent; and all the non-scary plans, can't save our skins!"

To which my Richard-model answers "But your concrete visualizations assume the endgame happens the day after tomorrow, at least politically. The future tends to go sideways! The endgame will likely happen in an environment quite different from our own! These day-after-tomorrow visualizations don't feel like they teach me much, because I think there's a good chance that the endgame-world looks dramatically different."

To which my Eliezer-model replies "Indeed, the future tends to go sideways. But I observe that the imagined changes, that I have heard so far, seem quite positive -- the relevant political actors become AI-savvy, the major states start coordinating, etc. I am quite suspicious of these sorts of visualizations, and would take them much more seriously if there was at least as much representation of outcomes as realistic as "then Trump becomes president" or "then at-home covid tests are banned in the US". And if all the ways to save the world *today* are scary-adjacent, the fact that the future is surprising gives us no *specific* reason to hope for that particular parameter to favorably change when the future in fact goes sideways. When things look grim, one can and should prepare to take advantage of miracles, but banking on some particular miracle is foolish."

And my Richard-model gets fuzzy at this point, but I'd personally be pretty enthusiastic about Richard naming a bunch of specific scenarios, not as predictions, but as the sorts of visualizations that seem to him promising, in the hopes of getting a much more object-level sense of why, in specific concrete scenarios, they either have the properties Eliezer is confident in, or are implausible on Eliezer's model (or surprise Eliezer and cause him to update).

[Tallinn][0:06] (Sep. 19)

excellent summary, nate! it also tracks my model of the debate well and summarises the frontier concisely (much better than your earlier notes or mine). unless eliezer or richard find major bugs in your summary, i'd nominate you to iterate after the next round of debate

[Soares: ❤]

7.3. Richard Ngo's summary

[Ngo][1:48] (Sep. 20)

Updated my summary to include the third discussion:

[https://docs.google.com/document/d/1sr5YchErvSAY2I4EkJl2dapHcMp8oCXy7g8hd_UajVw/edit]

I'm also halfway through a document giving my own account of intelligence + specific safe scenarios.

[Soares: ☺]

Why I'm excited about Redwood Research's current project

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

[Redwood Research's current project](#) is to train a model that completes short snippets of fiction without outputting text where someone gets injured. I'm excited about this direction and wanted to explain why.

(Disclaimer: I originally proposed this project, and am on Redwood's board. So I may be biased in favor of Redwood and especially in favor of this problem being a good one to work on.)

Relevance to deceptive alignment

I think about alignment risk as having two big drivers :

1. Your AI is smart enough that you couldn't even tell if it did something terrible until it's too late, e.g. until you've already implemented the proposed plan and it killed everyone.
2. Your AI looks nice during training while humans are in control, but there is a treacherous turn once it is deployed or becomes smart enough that we can't correct a problem.

I spend most of my time working on problem #1, but I think both problems are very important.

One of the main concrete candidate techniques for fixing problem #2 is to run the AI on tons of simulated inputs designed to *look like* the kind of test-time situation where your AI could kill you and take your stuff (adversarial training). But in fact these are just simulations, and if your model tries to do something really bad you just do a gradient update and move on. As long as you've solved problem #1, you can hopefully recognize if the model tries to do something bad in simulation without needing to run it in the real world.

Depending on the details you may not actually want to train on these examples---once you see a problem you may want to do something different or else run the risk of overfitting---but at any rate you definitely want to see them in order to know that there is a problem.)

I think adversarial training or evaluation look like key ingredients whether we are pursuing a principled approach to alignment (a more sophisticated version of this process are a key ingredient in my own "best guess" alignment plan) or just "winging it" (in which case I think adversarial evaluation is especially important).

Redwood's current project

I'm discussing what [Redwood's post](#) calls "step 1," namely training a classifier that is able to recognize violent completions with an extremely low false negative rate. Ideally they would go for 0 false negatives, but in practice they are just going for as many 9s as they can get (think more like 10^{-30} than 99.99%---way beyond what people are normally going for in ML). They are OK with a relatively high false positive rate since most inputs in practice are "obviously safe" and it's not that bad to force the model to behave cautiously on the riskiest inputs.

I think this overall structure is pretty similar to the long-term problem. We just need enough 9s that the model will probably *never* fail catastrophically even if you run it on many times more examples than it was trained on and even if the input distribution shifts in an unlucky way. And it's pretty OK if you have high false positive rates off distribution so that your system behaves conservatively until you retrain.

Step 1 is particularly hard because it's the place where we don't have a precise specification and instead need to use expensive oversight. I'm also interested in step 2---I think the two involve slightly different techniques but that both are quite relevant to the long-term problem. I'm genuinely unsure whether "learn a specification with low false negative rate" or "verify that your policy meets the learned specification" is the harder part.

"Try to find inputs where your classifier fails to recognize someone getting obviously injured" isn't a perfect analogy for "try to find inputs where your policy deliberately kills everyone," but I think it's actually surprisingly good. I think the core common features are that the task is extremely easy, so that the model *ought* to be able to do it without needing to learn very much new about the world, but the way ML systems are typically trained makes it hard to get good robustness in the tails.

Why the concrete steps towards this goal look interesting

I think that the concrete steps you would need to take to solve the problem today are usefully analogous to the steps you'll need to take in the future, and when I look at what Redwood folks are actually doing day-to-day I feel pretty happy about it:

- Organizing humans to create adversarial prompts where the model may do something bad, and figuring out how to combine this data with on-distribution data. Avoiding failures where the model learns to be conservative only on "sneaky-looking prompts designed to trick me."
- Writing tools to help improve that red-teaming process---using ML to automate parts of the task, and poking at your model to help guide the search process.
- Reasoning about the "margin of safety" between the worst example the red team can find, and the level of badness that would be needed to actually cause a serious problem. Thinking about how "worst we can find" relates to "actual worst," and how to trade this margin off against quality.
- Generating ensembles of classifiers and using different classifier strategies to get a variety of different answers to the question "was this behavior bad?" Selecting and synthesizing situations that create disagreements between these methods as a way of finding possible mistakes.

Many of those techniques appear in the academic literature, but the details change a lot when you start aiming for *no failures* rather than *reducing the rate of failures*.

I think that all of those will be relevant in the future. Human red-teaming will become less important relative to automated red-teaming, but I think that {human red-teaming with machine help} is more similar to what we'll be doing in the future than simpler automated techniques like gradient descent.

I don't think this is by any means a *complete* list of techniques that would be needed. But still, when I look at the actual activities and the project that puts them all together, I feel like this is the kind of work which I really want to be happening (in a massively scaled-up way) for models that are capable enough to be really risky.

Limitations of the analogy

There's one deep disanalogy between the current project and deceptive alignment: the current model is making "honest mistakes" when it completes a prompt in an injurious way, rather than biding its time until it can do something treacherous. That leads to two important differences:

- The failure cases aren't strategically hidden, so it's more likely that you can find them by searching kind of haphazardly. This problem will get much harder for the case of deceptive alignment. The current techniques still seem useful, and I think this is a good way to warm up, but even with more powerful automation I think they eventually won't be enough.
- It's not necessarily easy for the model to avoid failing, because it often doesn't understand when a situation involves an injury---it just says some plausible-sounding words that happen to imply that someone was injured. In practice this means that working on this problem involves some awkward judgment calls, where the model does say things that technically imply an injury occurred but where it can't connect the dots. I don't think this really affects the value of the research, but it makes it much harder to give a super clean evaluation of the outcome.

I think that these reduce the value of the project, but given the early stages it's more like 2x than 100x less useful. In many respects the current problem is like an easy version of the long-term problem, and some of the gap is closed by having adversaries working with ML tools that will scale up over time.

Why it matters

- I think it's plausible that evaluating and optimizing worst-case performance involves a bunch of skills and techniques rather than a unified idea, and that whether it succeeds or fails may depend on having teams that are really good at what they do rather than having some clever ideas (or may depend on both). So I want us to get as much practice as we can.
- I think that clever techniques for optimizing worst case performance will be easier to develop in the context of concrete problems where we've already picked the low-hanging fruit and have a clear understanding of why it's hard to make further progress without a clever idea. So even if none of the techniques involved in this project turn out to be valuable at the end of days, I still think it would be useful for laying the groundwork for further work.
- Existing work on worst-case performance is often focused on settings like classification accuracy given adversarial perturbations. But "classify well" seems

like too challenging a property to satisfy in the worst case---there's no clear reason why a model would be able to succeed, whereas I do think that it's possible to not deliberately kill everyone regardless of your input. Moreover, when these tasks are possible, it's often because the domain has some special structure (e.g. sparse or small perturbations) that a defender needs to exploit.

What I want to happen in the world overall

As we approach transformative and catastrophically risky AI, I expect there to be multiple industry labs working on training ever-larger models.

In my ideal world those labs would have large “adversarial evaluation departments” that try extremely hard to find inputs (or random seeds, or [“pseudo” inputs](#)) where a powerful model attempts to deliberately cause harm, or do anything that even vaguely smells like causing harm or deliberately undermining safety measures, or trying to deceptively hide their capabilities, or etc. We can allow the adversaries all kinds of unrealistic flexibility to create implausible situations, remove or relax all the realistic safety thresholds that might prevent bad behavior, and so on.

Ideally by the time a model is deployed we'd be pretty confident it's very hard to get it to do anything close to problematic. This won't be enough on its own to be confident that models don't do anything bad, and ideally this would be just one piece of a machine that created much more confidence, but it does make life meaningfully harder for a deceptively aligned model looking to cause trouble.

These teams would grow in size rapidly as AI became more dangerous, but I think things are much more likely to go well if people aren't trying to spin them up from scratch---if there are existing teams and individuals that are excellent at this activity, if they've been testing and developing methods for years, if we collectively understand what it looks like for such a team to be doing its job well, and so on.

Ideally that scale-up will be continuous, and will start up in something like the world of today, rather than being some hypothetical that we start executing on much closer to crunch time.

If all of that goes well, then the most likely way for us to fail is if the deceptively aligned model is a lot smarter than the red team (and all of the red team's aligned helpers)---so that the deceptive model can cause trouble in ways too subtle for the red team to notice. This basically puts us back to bullet point #1 in my original dichotomy. It's not that I'm not worried about that problem---I personally spend nearly 100% of my time on it---but I do think that it would be extremely comforting if that was *all* we had to worry about.

There are a lot of complicated details in that story. For example, it's not clear if you want to train on the data. And in practice you'd love to red-team the red-team by looking for cases where their methodology *fails* to catch what would be critical problems in synthetic cases. And for really powerful models there isn't going to be a clean development/deployment split---we will want to be red teaming checkpoints of the model as we go (and even then you may need to exert more care). And this entire rendition of the story is a bit of an unrealistic caricature in more pragmatic ways e.g. by positing a single dedicated team and so on.

Details aside, I think there is an unusually clear connection between this kind of work on characterizing/improving worst-case performance and a plausibly critical task for

reducing catastrophic risk from very capable models.

Christiano, Cotra, and Yudkowsky on AI progress

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

This post is a transcript of a discussion between Paul Christiano, Ajeya Cotra, and Eliezer Yudkowsky on AGI forecasting, following up on Paul and Eliezer's ["Takeoff Speeds" discussion](#).

Color key:

Chat by Paul and Eliezer Chat by Ajeya Inline comments

8. September 20 conversation

8.1. Chess and Evergrande

[Christiano][15:28]

I still feel like you are overestimating how big a jump alphago is, or something. Do you have a mental prediction of how the graph of (chess engine quality) vs (time) looks, and whether neural net value functions are a noticeable jump in that graph?

Like, people investing in "Better Software" doesn't predict that you won't be able to make progress at playing go. The reason you can make a lot of progress at go is that there was extremely little investment in playing better go.

So then your work is being done by the claim "People won't be working on the problem of acquiring a decisive strategic advantage," not that people won't be looking in quite the right place and that someone just had a cleverer idea

[Yudkowsky][16:35]

I think I'd expect something like... chess engine slope jumps a bit for Deep Blue, then levels off with increasing excitement, then jumps for the Alpha series? Albeit it's worth noting that Deepmind's effort there were going towards generality rather than raw power; chess was solved to the point of being uninteresting, so they tried to solve chess with simpler code that did more things. I don't think I do have strong opinions about what the chess trend should look like, vs. the Go trend; I have no memories of people saying the chess trend was breaking upwards or that there was a surprise there.

Incidentally, the highly well-traded financial markets are currently experiencing sharp dips surrounding the Chinese firm of Evergrande, which I was reading about several weeks before this.

I don't see the basic difference in the kind of reasoning that says "Surely foresighted firms must prod investments well in advance into earlier weaker applications of AGI that will double the economy", and the reasoning that says "Surely world economic markets and particular Chinese stocks should experience smooth declines as news about Evergrande becomes better-known and foresighted financial firms start to remove that stock from their portfolio or short-sell it", except that in the latter case there are many more actors with lower barriers to entry than presently exist in the auto industry or semiconductor industry never mind AI.

or if not smooth because of bandwagoning and rational fast actors, then at least the markets should (arguedo) be reacting earlier than they're reacting now, given that I heard about Evergrande earlier and they should have options-priced Covid earlier; and they should have reacted to the mortgage market earlier. If even markets there can exhibit seemingly late wild swings, how is the economic impact of AI - which isn't even an asset market! - forced to be earlier and smoother than that, as a result of investing?

There's just such a vast gap between hopeful reasoning about how various agents and actors should do the things the speaker finds very reasonable, thereby yielding smooth behavior of the Earth, versus reality.

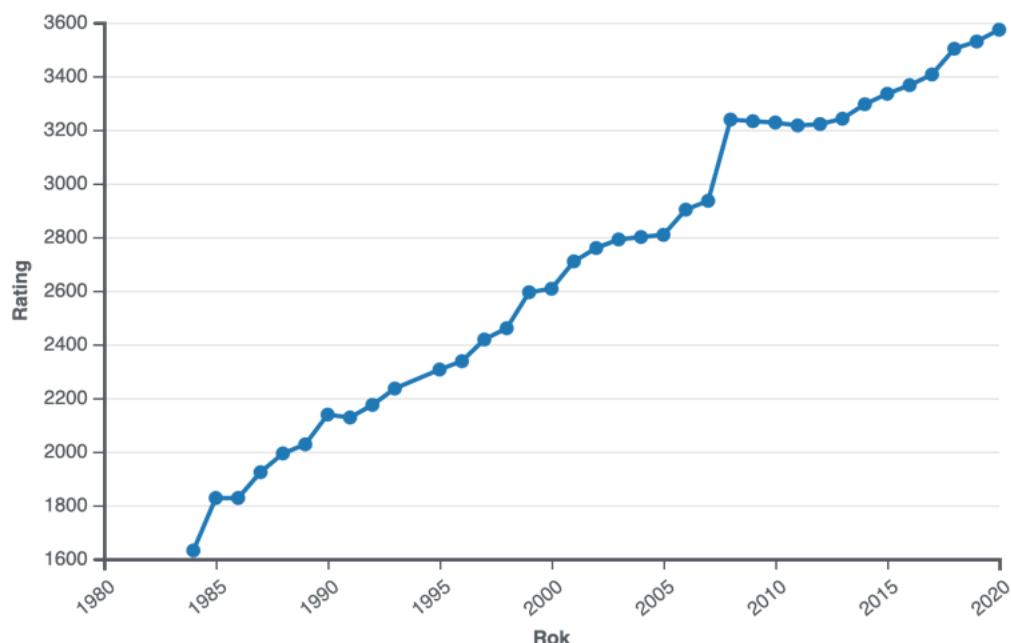
9. September 21 conversation

9.1. AlphaZero, innovation vs. industry, the Wright Flyer, and the Manhattan Project

[Christiano][10:18]

(For benefit of readers, the market is down 1.5% from friday close -> tuesday open, after having drifted down 2.5% over the preceding two weeks. Draw whatever lesson you want from that.)

Also for the benefit of readers, here is the SSDF list of computer chess performance by year. I think the last datapoint is with the first version of neural net evaluations, though I think to see the real impact want to add one more datapoint after the neural nets are refined (which is why I say I also don't know what the impact is)



No one keeps similarly detailed records for Go, and there is much less development effort, but the rate of progress was about 1 stone per year from 1980 until 2015 (see <https://intelligence.org/files/AlgorithmicProgress.pdf>, written way before AGZ). In 2012 go bots reached about 4-5 amateur dan. By DeepMind's reckoning here (<https://www.nature.com/articles/nature16961> figure 4) Fan AlphaGo about 4-5 stones stronger 4 years later, with 1 stone explained by greater compute. They could then get further progress to be superhuman with even more compute, radically more than were used for previous projects and with pretty predictable scaling. That level is within 1-2 stones of the best humans (professional dan are greatly compressed relative to amateur dan), so getting to "beats best human" is really just not a big discontinuity and the fact that DeepMind market can find an expert who makes a really bad forecast shouldn't be having such a huge impact on your view.

This understates the size of the jump from AlphaGo, because that was basically just the first version of the system that was superhuman and it was still progressing very rapidly as it moved from prototype to slightly-better-prototype, which is why you saw such a close game. (Though note that the AlphaGo prototype involved much more engineering effort than any previous attempt to play go, so it's not surprising that a "prototype" was the thing to win.)

So to look at actual progress after the dust settles and really measure how crazy this was, it seems much better to look at AlphaZero which continued to improve further, see (<https://sci-hub.se/https://www.nature.com/articles/nature24270>, figure 6b). Their best system got another ~8 stones of progress over AlphaGo. Now we are like 7-10 stones ahead of trend, of which I think about 5 stones are explained by compute. Maybe call it 6 years ahead of schedule?

So I do think this is pretty impressive, they were slightly ahead of schedule for beating the best human, but they did it with a huge margin of error. I think the margin is likely overstated a bit by their elo evaluation methodology, but I'd still grant like 5 years ahead of the nearest competition.

I'd be interested in input from anyone who knows more about the actual state of play (+ is allowed to talk about it) and could correct errors.

Mostly that whole thread is just clearing up my understanding of the empirical situation, probably we still have deep disagreements about what that says about the world, just as e.g. we read very different lessons from market movements.

Probably we should only be talking about either ML or about historical technologies with meaningful economic impacts. In my view your picture is just radically unlike how almost any technologies have been developed over the last few hundred years. So probably step 1 before having bets is to reconcile our views about historical technologies, and then maybe as a result of that we could actually have a better understanding about future technology. Or we could try to shore up the GDP bet.

Like, it feels to me like I'm saying: AI will be like early computers, or modern semiconductors, or airplanes, or rockets, or cars, or trains, or factories, or solar panels, or genome sequencing, or basically anything else. And you are saying: AI will be like nuclear weapons.

I think from your perspective it's more like: AI will be like all the historical technologies, and that means there will be a hard takeoff. The only way you get a soft takeoff forecast is by choosing a really weird thing to extrapolate from historical technologies.

So we're both just forecasting that AI will look kind of like other stuff in the near future, and then both taking what we see as the natural endpoint of that process.

To me it feels like the nuclear weapons case is the outer limit of what looks plausible, where someone would be able to spend \$100B for a chance at a decisive strategic advantage.

[Yudkowsky][11:11]

Go-wise, I'm a little concerned about that "stone" metric - what would the chess graph look like if it were measuring pawn handicaps? Are the professional dans compressed in Elo, not just "stone handicaps" relative to the amateur dans? And I'm also kinda surprised by the claim, which I haven't yet looked at, that Alpha Zero got 8 stones of progress over AlphaGo - I would not have been shocked if you told me that God's Algorithm couldn't beat Lee Se-dol with a 9-stone handicap.

Like, the obvious metric is Elo, so if you go back and refigure in "stone handicaps", an obvious concern is that somebody was able to look into the past and fiddle their hindsight until they found a hindsight metric that made things look predictable again. My sense of Go said that 5-dan amateur to 9-dan pro

was a HELL of a leap for 4 years, and I also have some doubt about the original 5-dan-amateur claim and whether those required relatively narrow terms of testing (eg timed matches or something).

One basic point seems to be whether AGI is more like an innovation or like a performance metric over entire large industry.

Another point seems to be whether the behavior of the world is usually like that, in some sense, or if just that people who like smooth graphs can go find some industries that have smooth graphs for particular performance metrics that happen to be smooth.

Among the smoothest metrics I know that seems like a convergent rather than handpicked thing to consider is world GDP, which is the sum of more little things than almost anything else, and whose underlying process is full of multiple stages of converging-product-line bottlenecks that make it hard to jump the entire GDP significantly even when you jump one component of a production cycle... which, from my standpoint, is a major reason to expect AI to not hit world GDP all that hard until AGI passes the critical threshold of bypassing it entirely. Having 95% of the tech to invent a self-replicating organism (eg artificial bacterium) does not get you 95%, 50%, or even 10% of the impact.

(it's not so much the 2% reaction of world markets to Evergrande that I was singling out earlier, 2% is noise-ish, but the wider swings in the vicinity of Evergrande particularly)

[Christiano][12:41]

Yeah, I'm just using "stone" to mean "elo difference that is equal to 1 stone at amateur dan / low kyu" you can see DeepMind's conversion (which I also don't totally believe) in figure 4 here (<https://sci-hub.se/https://www.nature.com/articles/nature16961>). Stones are closer to constant elo than constant handicap, it's just a convention to name them that way.

[Yudkowsky][12:42]

k then

[Christiano][12:47]

But my description above still kind of understates the gap I think. They call 230 elo 1 stone, and I think the prior rate of progress is more like 200 elo/year. They put AlphaZero about 3200 elo above the 2012 Go system, so that's like 16 years ahead = 11 years ahead of schedule. At least 2 years are from test-time hardware, and self-play systematically overestimates elo differences at the upper end of that. But 5 years ahead is still too low and that sounds more like 7-9 years ahead. ETA: and my actual best guess for when things considered is probably 10 years ahead, which I agree is just a lot bigger than 5. And I also understated how much of the gap was getting up to Lee Sedol.

The go graph I posted wasn't made with hindsight, that was from 2014

I mean, I'm fine with you saying that people who like smooth graphs are cherry-picking evidence, but you want to give any example other than nuclear weapons of technologies with the kind of discontinuous impact you are describing?

I do agree that the difference in our views is like "innovation" vs "industry." And a big part of my position is that innovation-like things just don't usually have big impacts for kind of obvious reasons, they start small and then become more industry-like as they scale up. And current deep learning seems like an absolutely stereotypical industry that is scaling up rapidly in an increasingly predictable way.

As far as I can tell the examples we know of things changing continuously aren't handpicked, we've been looking at all the examples we can find, and no one is proposing or even able to find almost *anything* that looks like you are imagining AI will look.

Like, we've seen deep learning innovations in the form of prototypes (most of all AlexNet), and they were cool and represented giant fast changes in people's views. And more recently we are seeing big much-less-surprising changes that are still helping a lot in raising the tens of billions of dollars that people are raising. And the innovations we are seeing are increasingly things that trade off against modest improvements in model size, there are fewer and fewer big surprises, just like you'd predict.

clearer and clearer to more and more people what the roadmap is---the roadmap is not yet quite as clear as in semiconductors, but as far as I can tell that's just because the field is still smaller.

[Yudkowsky][13:23]

I sure wasn't imagining there was a roadmap to AGI! Do you perchance have one which says that AG 30 years out?

From my perspective, you could as easily point to the Wright Flyer as an atomic bomb. Perhaps this reflects again the "innovation vs industry" difference, where I think in terms of building a thing that grows thereby bypassing our small cute world GDP, and you think in terms of industries that affect whole GDP in an invariant way throughout their lifetimes.

Would you perhaps care to write off the atomic bomb too? It arguably didn't change the outcome of World War II or do much that conventional weapons in great quantity couldn't; Japan was bluffed into believing the US could drop a nuclear bomb every week, rather than the US actually having that many nuclear bombs or them actually being used to deliver a historically outsized impact on Japan. From the industry-centric perspective, there is surely some graph you can draw which makes nuclear weapons also look like business as usual, especially if you go by destruction per unit of whole-industry non-marginal expense, rather than destruction per bomb.

[Christiano][13:27]

seems like you have to make the Wright Flyer much better before it's important, and that it becomes more like an industry as that happens, and that this is intimately related to why so few people were working on it

I think the atomic bomb is further on the spectrum than almost anything, but it still doesn't feel near as far as what you are expecting out of AI

the Manhattan Project took years and tens of billions; if you wait an additional few years and spend an additional few tens of billions then it would be a significant improvement in destruction or deterrence per \$ (but not totally insane)

I do think it's extremely non-coincidental that the atomic bomb was developed in a country that was practically outspending the whole rest of the world in "killing people technology"

and took a large fraction of that country's killing-people resources

eh, that's a bit unfair, the US was only like 35% of global spending on munitions

and the Manhattan Project itself was only a couple percent of total munitions spending

[Yudkowsky][13:32]

a lot of why I expect AGI to be a disaster is that *I am straight-up expecting AGI to be different*. If it were just like coal or just like nuclear weapons or just like viral biology then I would not be way more worried about AGI than I am worried about those other things.

[Christiano][13:33]

that definitely sounds right

but it doesn't seem like you have any short-term predictions about AI being different

9.2. AI alignment vs. biosafety, and measuring progress

[Yudkowsky][13:33]

are you more worried about AI than about bioengineering?

[Christiano][13:33]

I'm more worried about AI because (i) alignment is a thing, unrelated to takeoff speed, (ii) AI is a (ET/ likely to be) huge deal and bioengineering is probably a relatively small deal

(in the sense of e.g. how much \$ people spend, or how much \$ it makes, or whatever other metric of size you want to use)

[Yudkowsky][13:35]

what's the disanalogy to (i) biosafety is a thing, unrelated to the speed of bioengineering? why expect AI to be a huge deal and bioengineering to be a small deal? is it just that investing in AI is scaling faster than investment in bioengineering?

[Christiano][13:35]

no, alignment is a really easy x-risk story, bioengineering x-risk seems extraordinarily hard

It's really easy to mess with the future by creating new competitors with different goals, if you want to mess with the future by totally wiping out life you have to really try at it and there's a million ways it can fail. The bioengineering seems like it basically requires deliberate and reasonably competent malice whereas alignment seems like it can only be averted with deliberate effort, etc.

I'm mostly asking about historical technologies to try to clarify expectations, I'm pretty happy if the outcome is: you think AGI is predictably different from previous technologies in ways we haven't seen yet

though I really wish that would translate into some before-end-of-days prediction about a way that AGI will eventually look different

[Yudkowsky][13:38]

in my ontology a whole lot of threat would trace back to "AI hits harder, faster, gets too strong to be adjusted"; tricks with proteins just don't have the raw power of intelligence

[Christiano][13:39]

in my view it's nearly totally orthogonal to takeoff speed, though fast takeoffs are a big reason that preparation in advance is more useful

(but not related to the basic reason that alignment is unprecedently scary)

It feels to me like you are saying that the AI-improving-AI will move very quickly from "way slower than humans" to "FOOM in <1 year," but it just looks like that is very surprising to me.

However I do agree that if AI-improving-AI was like AlphaZero, then it would happen extremely fast.

It seems to me like it's pretty rare to have these big jumps, and it gets much much rarer as technologies become more important and are more industry-like rather than innovation like (and people care about them a lot rather than random individuals working on them, etc.). And I can't tell whether you are saying something more like "nah big jumps happen all the time in places that are structurally analogous to the key takeoff jump, even if the effects are blunted by slow adoption and regulatory bottlenecks and so on" or if you are saying "AGI is atypical in how jumpy it will be"

[Yudkowsky][13:44]

I don't know about *slower*; GPT-3 may be able to type faster than a human

[Christiano][13:45]

Yeah, I guess we've discussed how you don't like the abstraction of "speed of making progress"

[Yudkowsky][13:45]

but, basically less useful in fundamental ways than a human civilization, because they are less complete, less self-contained

[Christiano][13:46]

Even if we just assume that your AI needs to go off in the corner and not interact with humans, there still a question of why the self-contained AI civilization is making ~0 progress and then all of a sudden very rapid progress

[Yudkowsky][13:46]

unfortunately a lot of what you are saying, from my perspective, has the flavor of, "but can't you tell about your predictions earlier on of the impact on global warming at the *Homo erectus* level"

you have stories about why this is like totally not a fair comparison

I do not share these stories

[Christiano][13:46]

I don't understand either your objection nor the reductio

like, here's how I think it works: AI systems improve gradually, including on metrics like "How long does it take them to do task X?" or "How high-quality is their output on task X?"

[Yudkowsky][13:47]

I feel like the thing we know is something like, there is a sufficiently high level where things go whooo... humans-from-hominids style

[Christiano][13:47]

We can measure the performance of AI on tasks like "Make further AI progress, without human input"

Any way I can slice the analogy, it looks like AI will get continuously better at that task

[Yudkowsky][13:48]

how would you measure progress from GPT-2 to GPT-3, and would you feel those metrics really capture the sort of qualitative change that lots of people said they felt?

[Christiano][13:48]

And it seems like we have a bunch of sources of data we can use about how fast AI will get better

Could we talk about some application of GPT-2 or GPT-3?

also that's a *lot* of progress, spending 100x more is a *lot* more money

[Yudkowsky][13:49]

my world, GPT-3 has very few applications because it is not quite right and not quite complete

[Christiano][13:49]

also it's still really dumb

[Yudkowsky][13:49]

like a self-driving car that does great at 99% of the road situations

economically almost worthless

[Christiano][13:49]

I think the "being dumb" is way more important than "covers every case"

[Yudkowsky][13:50]

(albeit that if new cities could still be built, we could totally take those 99%-complete AI cars and build fences and fence-gates around them, in a city where they were the only cars on the road, in which case they *would* work, and get big economic gains from these new cities with driverless cars, which ties back into my point about how current world GDP is *unwilling* to accept tech inputs)

like, it is in fact very plausible to me that there is a neighboring branch of reality with open borders and no housing-supply-constriction laws and no medical-supply-constriction laws, and their world GDP does manage to double before AGI hits them really hard, albeit maybe not in 4 years. this world is not Earth, they are constructing new cities to take advantage of 99%-complete driverless cars *right now*, or rather, they started constructing them 5 years ago and finished 4 years and 6 months ago.

9.3. Requirements for FOOM

[Christiano][13:53]

I really feel like the important part is the jumpiness you are imagining on the AI side / why AGI is different from other things

[Cotra][13:53]

It's actually not obvious to me that Eliezer is imagining that much more jumpiness on the AI technology side than you are, Paul

E.g. he's said in the past that while the gap from "subhuman to superhuman AI" could be 2h if it's in the middle of FOOM, it could also be a couple years if it's more like scaling alphago

[Yudkowsky][13:54]

Indeed! We observed this jumpiness with hominids. A lot of stuff happened at once with hominids, because a critical terminal part of the jump was the way that hominids started scaling their own food supply, instead of being ultimately limited by the food supply of the savanna.

[Cotra][13:54]

A couple years is basically what Paul believes

[Christiano][13:55]

(discord is not a great place for threaded conversations :()

[Cotra][13:55]

What are the probabilities you're each placing on the 2h-2y spectrum? I feel like Paul is like "no way < 2h, likely on 2y" and Eliezer is like "who knows" on the whole spectrum, and a lot of the disagreement is about the impact of the previous systems?

[Christiano][13:55]

yeah, I'm basically at "no way," because it seems obvious that the AI that can foom in 2h is preceded by the AI that can foom in 2y

[Yudkowsky][13:56]

well, we surely agree there!

[Christiano][13:56]

OK, and it seems to me like it is preceded by years

[Yudkowsky][13:56]

we disagree on whether the AI that can foom in 2y clearly comes more than 2y before the AI that fooms in 2h

[Christiano][13:56]

yeah

perhaps we can all agree it's preceded by at least 2h

so I have some view like: for any given AI we can measure "how long does it take to foom?" and it seems to me like this is just a nice graph

and it's not exactly clear how quickly that number is going down, but a natural guess to me is something like "halving each year" based on the current rate of progress in hardware and software

and you see localized fast progress most often in places where there hasn't yet been much attention

and my best guess for your view is that actually that's not a nice graph at all, there is some critical threshold or range where AI quickly moves from "not foaming for a really long time" to "foaming real fast," and that seems like the part I'm objecting to

[Cotra][13:59]

Paul, is your take that there's a non-infinity number for time to FOOM that'd be associated with current AI systems (unassisted by humans)?

And it's going down over time?

I feel like I would have said something more like "there's a \$ amount it takes to build a system that will FOOM in X amount of time, and that's going down"

where it's like quadrillions of dollars today

[Christiano][14:00]

I think it would be a big engineering project to make such an AI, which no one is doing because it would be uselessly slow even if successful

[Yudkowsky][14:02]

I... don't think GPT-3 fooms given 2^{30} longer time to think about than the systems that would otherwise exist 30 years from now, on timelines I'd consider relatively long, and hence generous to the viewpoint? I also don't think you can take a quadrillion dollars and scale GPT-3 to foom today?

[Cotra][14:03]

I would agree with your take on GPT-3 fooming, and I didn't mean a quadrillion dollars just to scale GPT-3, would probably be a diff't architecture

[Christiano][14:03]

I also agree that GPT-3 doesn't foom, it just keeps outputting <EOT>[next web page]<EOT>...

But I think the axes of "smart enough to foom fast" and "wants to foom" are pretty different. I also agree there is some minimal threshold below which it doesn't even make sense to talk about "wants to foom", which I think is probably just not that hard to reach.

(Also there are always diminishing returns as you continue increasing compute, which become very relevant if you try to GPT-3 for a billion billion years as in your hypothetical even apart from "wants to foom".)

[Cotra][14:06]

I think maybe you and EY then disagree on where the threshold from "infinity" to "a finite number" for "time for this AI system to FOOM" begins? where eliezer thinks it'll drop from infinity to a pretty small finite number and you think it'll drop to a pretty large finite number, and keep going down from there

[Christiano][14:07]

I also think we will likely jump down to a foom-ing system only after stuff is pretty crazy, but I think that's probably less important

I think what you said is probably the main important disagreement

[Cotra][14:08]

as in before that point it'll be faster to have human-driven progress than FOOM-driven progress bc the FOOM would be too slow?

and there's some crossover point around when the FOOM time is just a bit faster than the human-driven progress time

[Christiano][14:09]

yeah, I think most likely (AI+humans) is faster than (AI alone) because of complementarity. But I thin Eliezer and I would still disagree even if I thought there was 0 complementarity and it's just (humans improving AI) and separately (AI improving AI)

on that pure substitutes model I expect "AI foom" to start when the rate of AI-driven AI progress overtakes the previous rate of human-driven AI progress

like, I expect the time for successive "doublings" of AI output to be like 1 year, 1 year, 1 year, 1 year, takes over] 6 months, 3 months, ...

and the most extreme fast takeoff scenario that seems plausible is that kind of perfect substitutes + physical economic impact from the prior AI systems

and then by that point fast enough physical impact is really hard so it happens essentially after the software-only singularity

I consider that view kind of unlikely but at least coherent

9.4. AI-driven accelerating economic growth

[Yudkowsky][14:12]

I'm expecting that the economy doesn't accept much inputs from chimps, and then the economy doesn't accept much input from village idiots, and then the economy doesn't accept much input from weird immigrants. I can imagine that there may or may not be a very weird 2-year or 3-month period with strange half-genius systems running around, but they will still not be allowed to build houses. In the terminal phase things get more predictable and the AGI starts its own economy instead.

[Christiano][14:12]

I guess you can go even faster, by having a big and accelerating ramp-up in human investment right around the end, so that the "1 year" is faster (e.g. if recursive self-improvement was like playing go, you could move from "a few individuals" to "google spending \$10B" over a few years)

[Yudkowsky][14:13]

My model prophecy doesn't rule that out as a thing that could happen, but sure doesn't emphasize it as a key step that needs to happen.

[Christiano][14:13]

I think it's very likely that AI will mostly be applied to further hardware+software progress

[Cotra: +]

I don't really understand why you keep talking about houses and healthcare

[Cotra][14:13]

Eliezer, what about stuff like Google already using ML systems to automate its TPU load-sharing decisions, and people starting to use Codex to automate routine programming, and so on? Seems like there's a lot of stuff like that starting to already happen and markets are pricing in huge further increases

[Christiano][14:14]

it seems like the non-AI up-for-grabs zone are things like manufacturing, not things like healthcare

[Cotra: +]

[Cotra][14:14]

(I mean on your timelines obviously not much time for acceleration anyway, but that's distinct from t regulation not allowing weak AIs to do stuff story)

[Yudkowsky][14:14]

Because I think that a key thing of what makes your prophecy less likely is the way that it happens inside the real world, where, economic gains or not, the System is unwilling/unable to take the things that are 99% self-driving cars and start to derive big economic benefits from those.

[Cotra][14:15]

but it seems like huge economic gains could happen entirely in industries mostly not regulated and n customer-facing, like hardware/software R&D, manufacturing, shipping logistics, etc

[Yudkowsky][14:15]

Ajeya, I'd consider Codex of *far* greater could-be-economically-important-ness than automated TPU load sharing decisions

[Cotra][14:15]

i would agree with that, it's smarter and more general

and i think that kind of thing could be applied on the hardware chip design side too

[Yudkowsky][14:16]

no, because the TPU load-sharing stuff has an obvious saturation point as a world economic input, w superCodex could be a world economic input in many more places

[Cotra][14:16]

the TPU load sharing thing was not a claim that this application could scale up to crazy impacts, but t it was allowed to happen, and future stuff that improves that kind of thing (back-end hardware/software/logistics) would probably also be allowed

[Yudkowsky][14:16]

my sense is that decuplicating the number of programmers would not lift world GDP much, but it seems lot more possible for me to be wrong about that

[Christiano][14:17]

the point is that housing and healthcare are not central examples of things that scale up at the beginning of explosive growth, regardless of whether it's hard or soft

they are slower and harder, and also in efficient markets-land they become way less important during the transition

so they aren't happening that much on anyone's story

and also it doesn't make that much difference whether they happen, because they have pretty limited effects on other stuff

like, right now we have an industry of ~hundreds of billions that is producing computing hardware, building datacenters, mining raw inputs, building factories to build computing hardware, solar panels shipping around all of those parts, etc. etc.

I'm kind of interested in the question of whether all that stuff explodes, although it doesn't feel as cool as the question of "what are the dynamics of the software-only singularity and how much \$ are people spending initiating it?"

but I'm not really interested in the question of whether human welfare is spiking during the transition only after

[Yudkowsky][14:20]

All of world GDP has never felt particularly relevant to me on that score, since twice as much hardware maybe corresponds to being 3 months earlier, or something like that.

[Christiano][14:21]

that sounds like the stuff of predictions?

[Yudkowsky][14:21]

But if complete chip manufacturing cycles have accepted much more effective AI input, with no non-bottlenecks, then that... sure is a much more *material* element of a foom cycle than I usually envision

[Christiano][14:21]

like, do you think it's often the case that 3 months of software progress = doubling compute spending or do you think AGI is different from "normal" AI on this perspective?

I don't think that's that far off anyway

I would guess like ~1 year

[Yudkowsky][14:22]

Like, world GDP that goes up by only 10%, but that's because producing compute capacity was 2.5% of world GDP and that quadrupled, starts to feel much more to me like it's part of a foom story.

I expect software-beats-hardware to hit harder and harder as you get closer to AGI, yeah.

the prediction is firmer near the terminal phase, but I think this is also a case where I expect that to be visible earlier

[Christiano][14:24]

I think that by the time that the AI-improving-AI takes over, it's likely that hardware+software manufacturing+R&D represents like 10-20% of GDP, and that the "alien accountants" visiting earth would value those companies at like 80%+ of GDP

9.5. Brain size and evolutionary history

[Cotra][14:24]

On software beating hardware, how much of your view is dependent on your belief that the chimp -> human transition was probably not mainly about brain size because if it were about brain size it would have happened faster? My understanding is that you think the main change is a small software innovation which increased returns to having a bigger brain. If you changed your mind and thought the chimp -> human transition was probably mostly about raw brain size, what (if anything) about your AI takeoff views would change?

[Yudkowsky][14:25]

I think that's a pretty different world in a lot of ways!

but yes it hits AI takeoff views too

[Christiano][14:25]

regarding software vs hardware, here is an example of asking this question for imangenet classification ("how much compute to train a model to do the task?"), with a bit over 1 year doubling times (<https://openai.com/blog/ai-and-efficiency/>). I guess my view is that we can make a similar graph for "compute required to make your AI FOOM" and that it will be falling significantly slower than 2x/year. And my prediction for other tasks is that the analogous graphs will also tend to be falling slower than 2x/year.

[Yudkowsky][14:26]

to the extent that I modeled hominid evolution as having been "dutifully schlep more of the same stuff predictably more of the same returns" that would correspond to a world in which intelligence was less scary, different, dangerous-by-default

[Cotra][14:27]

thanks, that's helpful. I looked around in [IEM](#) and other places for a calculation of how quickly we should have evolved to humans if it were mainly about brain size, but I only found qualitative statements. If there's a calculation somewhere I would appreciate a pointer to it, because currently it seems to me a story like "selection pressure toward general intelligence was weak-to-moderate because it wasn't actually *that* important for fitness, and this degree of selection pressure is consistent with brain size being the main deal and just taking a few million years to happen" is very plausible

[Yudkowsky][14:29]

well, for one thing, the prefrontal cortex expanded twice as fast as the rest

and iirc there's evidence of a lot of recent genetic adaptation... though I'm not as sure you could pinpoint it as being about brain-stuff or that the brain-stuff was about cognition rather than rapidly shifting motivations or something.

elephant brains are 3-4 times larger by weight than human brains (just looked up)

if it's that easy to get returns on scaling, seems like it shouldn't have taken that long for evolution to get there

[Cotra][14:31]

but they have fewer synapses (would compute to less FLOP/s by the standard conversion)

how long do you think it should have taken?

[Yudkowsky][14:31]

early dinosaurs should've hopped onto the predictable returns train

[Cotra][14:31]

is there a calculation?

you said in IEM that evolution increases organ sizes quickly but there wasn't a citation to easily follow on there

[Yudkowsky][14:33]

I mean, you could produce a graph of smooth fitness returns to intelligence, smooth cognitive returns: brain size/activity, linear metabolic costs for brain activity, fit that to humans and hominids, then show that obviously if hominids went down that pathway, large dinosaurs should've gone down it first because they had larger bodies and the relative metabolic costs of increased intelligence would've been lower at every point along the way

I do not have a citation for that ready, if I'd known at the time you'd want one I'd have asked Luke M it while he still worked at MIRI 😊

[Cotra][14:35]

cool thanks, will think about the dinosaur thing (my first reaction is that this should depend on the actual fitness benefits to general intelligence which might have been modest)

[Yudkowsky][14:35]

I suspect we're getting off Paul's crux, though

[Cotra][14:35]

yeah we can go back to that convo (though I think Paul would also disagree about this thing, and believes that the chimp to human thing was mostly about size)

sorry for hijacking

[Yudkowsky][14:36]

well, if at some point I can produce a major shift in EA viewpoints by coming up with evidence for a bunch of non-brain-size brain selection going on over those timescales, like brain-related genes where we can figure out how old the mutation is, I'd then put a lot more priority on digging up a paper like that

I'd consider it sufficiently odd to imagine hominids->humans as being primarily about brain size, given the evidence we have, that I do not believe this is Paul's position until Paul tells me so

[Christiano][14:49]

I would guess it's primarily about brain size / neuron count / cortical neuron count

and that the change in rate does mostly go through changing niche, where both primates and birds have this cycle of rapidly accelerating brain size increases that aren't really observed in other animals
it seems like brain size is increasing extremely quickly on both of those lines

[Yudkowsky][14:50]

why aren't elephants GI?

[Christiano][14:51]

mostly they have big brains to operate big bodies, and also my position obviously does not imply (big brain) ===(necessarily implies)==> general intelligence

[Yudkowsky][14:52]

I don't understand, in general, how your general position manages to strongly imply a bunch of stuff about AGI and not strongly imply similar stuff about a bunch of other stuff that sure sounds similar to me

[Christiano][14:52]

don't elephants have very few synapses relative to humans?

[Cotra: +]

how does the scale hypothesis possibly take a strong stand on synapses vs neurons? I agree that it takes a modest predictive hit from "why aren't the big animals much smarter?"

[Yudkowsky][14:53]

if adding more synapses just scales, elephants should be able to pay hominid brain costs for a much smaller added fraction of metabolism and also not pay the huge death-in-childbirth head-size tax because their brains and heads are already 4x as huge as they need to be for GI and now they just need some synapses, which are a much tinier fraction of their total metabolic cost.

[Christiano][14:54]

I mean, you can also make smaller and cheaper synapses as evidenced by birds
I'm not sure I understand what you are saying
it's clear that you can't say "X is possible metabolically, so evolution would do it"
or else you are confused about why primate brains are so bad

[Yudkowsky][14:54]

great, then smaller and cheaper synapses should've scaled many eons earlier and taken over the wo

[Christiano][14:55]

this isn't about general intelligence, this is a reductio of your position...

[Yudkowsky][14:55]

and here I had thought it was a reductio of your position...

[Christiano][14:55]

indeed

like, we all grant that it's metabolically possible to have small smart brains
and evolution doesn't do it
and I'm saying that it's also possible to have small smart brains
and that scaling brains up matters a lot

[Yudkowsky][14:56]

no, you grant that it's metabolically possible to have cheap brains full of synapses, which are therefo
on your position, smart

[Christiano][14:56]

birds are just smart

we know they are smart

this isn't some kind of weird conjecture

like, we can debate whether they are a "general" intelligence, but it makes no difference to this
discussion

the point is that they do more with less metabolic cost

[Yudkowsky][14:57]

on my position, the brain needs to invent the equivalents of ReLUs and Transformers and really rath
lot of other stuff because it can't afford nearly that many GPUs, and then the marginal returns on ad
expensive huge brains and synapses have increased enough that hominids start to slide down the
resulting fitness slope, which isn't even paying off in guns and rockets yet, they're just getting that
much intelligence out of it once the brain software has been selected to scale that well

[Christiano][14:57]

but all of the primates and birds have brain sizes scaling much faster than the other animals

like, the relevant "things started to scale" threshold is way before chimps vs humans

isn't it?

[Cotra][14:58]

to clarify, my understanding is that paul's position is "Intelligence is mainly about synapse/neuron co
and evolution doesn't care that much about intelligence; it cared more for birds and primates, and be
lines are getting smarter+bigger-brained." And eliezer's position is that "evolution should care a ton
about intelligence in most niches, so if it were mostly about brain size then it should have gone up to
human brain sizes with the dinosaurs"

[Christiano][14:58]

or like, what is the evidence you think is explained by the threshold being between chimps and huma

[Yudkowsky][14:58]

if hominids have less efficient brains than birds, on this theory, it's because (post facto handwave) birds are tiny, so whatever cognitive fitness gradients they face, will tend to get paid more in software and biological efficiency and biologically efficient software, and less paid in Stack More Neurons (even compared to hominids)

elephants just don't have the base software to benefit much from scaling synapses even though they are relatively cheaper for elephants

[Christiano][14:59]

@ajeya I think that intelligence is about a lot of things, but that size (or maybe "more of the same" changes that had been happening recently amongst primates) is the big difference between chimps and humans

[Cotra: 

[Cotra][14:59]

got it yeah i was focusing on chimp-human gap when i said "intelligence" there but good to be careful

[Yudkowsky][14:59]

I have not actually succeeded in understanding Why On Earth Anybody Would Think That If Not For That Really Weird Prior I Don't Get Either

re: the "more of the same" theory of humans

[Cotra][15:00]

do you endorse my characterization of your position above? "evolution should care a ton about intelligence in most niches, so if it were mostly about brain size then it should have gone up to human brain sizes with the dinosaurs"

in which case the disagreement is about how much evolution should care about intelligence in the dinosaur niche, vs other things it could put its skill points into?

[Christiano][15:01]

Eliezer, it seems like chimps are insanely smart compared to other animals, basically as smart as the get

so it's natural to think that the main things that make humans unique are also present in chimps

or at least, there was something going on in chimps that is exceptional

and should be causally upstream of the uniqueness of humans too

otherwise you have too many coincidences on your hands

[Yudkowsky][15:02]

ajeya: no, I'd characterize that as "the human environmental niche per se does not seem super-specific enough to be unique on a geological timescale, the cognitive part of the niche derives from increased cognitive abilities in the first place and so can't be used to explain where they got started, dinosaurs were larger than humans and would pay lower relative metabolic costs for added brain size and it is not the case that every species as large as humans was in an environment where they would not have benefited as much from a fixed increment of intelligence, hominids are probably distinguished from dinosaurs in having better neural algorithms that arose over intervening evolutionary time and there were better returns in intelligence on synapses that are more costly to humans than to elephants or large dinosaurs"

[Christiano][15:03]

I don't understand how you can think that hominids are the special step relative to something earlier or like, I can see how it's consistent, but I don't see what evidence or argument supports it
it seems like the short evolutionary time, and the fact that you also have to explain the exceptional qualities of other primates, cut extremely strongly against it

[Yudkowsky][15:04]

paul: indeed, the fact that dinosaurs didn't see their brain sizes and intelligences ballooning, says there must be a lot of stuff hominids had that dinosaurs didn't, explaining why hominids got much higher returns on intelligence per synapse. natural selection is enough of a smooth process that 95% of this stuff should've been in the last common ancestor of humans and chimps.

[Christiano][15:05]

it seems like brain size basically just increases faster in the smarter animals? though I mostly just know about birds and primates

[Yudkowsky][15:05]

that is what you'd predict from smartness being about algorithms!

[Christiano][15:05]

and it accelerates further and further within both lines

it's what you'd expect if smartness is about algorithms *and chimps and birds have good algorithms*

[Yudkowsky][15:06]

if smartness was about brain size, smartness and brain size would increase faster in the *larger animals* or the ones whose successful members *ate more food per day*

well, sure, I do model that birds have better algorithms than dinosaurs

[Cotra][15:07]

it seems like you've given arguments for "there was algorithmic innovation between dinosaurs and humans" but not yet arguments for "there was major algorithmic innovation between chimps and humans"?

[Christiano][15:08]

(much less that the algorithmic changes were not just more-of-the-same)

[Yudkowsky][15:08]

oh, that's *not* mandated by the model the same way. (between LCA of chimps and humans)

[Christiano][15:08]

isn't that exactly what we are discussing?

[Yudkowsky][15:09]

...I hadn't thought so, no.

[Cotra][15:09]

original q was:

On software beating hardware, how much of your view is dependent on your belief that the chimp > human transition was probably not mainly about brain size because if it were about brain size it would have happened faster? My understanding is that you think the main change is a small software innovation which increased returns to having a bigger brain. If you changed your mind and thought that the chimp -> human transition was probably mostly about raw brain size, what (if anything) about your AI takeoff views would change?

so i thought we were talking about if there's a cool innovation from chimp->human?

[Yudkowsky][15:10]

I can see how this would have been the more obvious intended interpretation on your viewpoint, and apologize

[Christiano][15:10]

(though i think paul would also disagree about this thing, and believes that the chimp to human thing was mostly about size)

Is what I was responding to in part

I am open to saying that I'm conflating size and "algorithmic improvements that are closely correlate with size in practice and are similar to the prior algorithmic improvements amongst primates"

[Yudkowsky][15:11]

from my perspective, the question is "how did that hominid->human transition happen, as opposed to there being an elephant->smartelephant or dinosaur->smartdinosaur transition"?

I expect there were substantial numbers of brain algorithm stuffs going on during this time, however because I don't think that synapses scale that well *with* the baseline hominid boost

[Christiano][15:11]

FWIW, it seems quite likely to me that there would be an elephant->smartelephant transition within tens of millions or maybe 100M years, and a dinosaur->smartdinosaur transition in hundreds of millions of years

and those are just cut off by the fastest lines getting there first

[Yudkowsky][15:12]

which I think does circle back to that point? actually I think my memory glitched and forgot the original point while being about this subpoint and I probably did interpret the original point as intended.

[Christiano][15:12]

namely primates beating out birds by a hair

[Yudkowsky][15:12]

that sounds like a viewpoint which would also think it much more likely that GPT-3 would foom in a billion years

where maybe you think that's unlikely, but I still get the impression your "unlikely" is, like, 5 orders o magnitude likelier than mine before applying overconfidence adjustments against extreme probabilit on both sides

yeah, I think I need to back up

[Cotra][15:15]

Is your position something like "at some point after dinosaurs, there was an algorithmic innovation th increased returns to brain size, which meant that the birds and the humans see their brains increasir quickly while the dinosaurs didn't"?

[Christiano][15:15]

it also seems to me like the chimp->human difference is in basically the same ballpark of the effect c brain size within humans, given modest adaptations for culture

which seems like a relevant sanity-check that made me take the "mostly hardware" view more seriou

[Yudkowsky][15:15]

there's a part of my model which very strongly says that hominids scaled better than elephants and that's why "hominids->humans but not elephants->superelephants"

[Christiano][15:15]

previously I had assumed that analysis would show that chimps were obviously way dumber than an extrapolation of humans

[Yudkowsky][15:16]

there's another part of my model which says "and it still didn't scale that well without algorithms, so should expect a lot of alleles affecting brain circuitry which rose to fixation over the period when hominid brains were expanding"

this part is strong and I think echoes back to AGI stuff, but it is not as *strong* as the much *more* overdetermined position that hominids started with more scalable algorithms than dinosaurs.

[Christiano][15:17]

I do agree with the point that there are structural changes in brains as you scale them up, and this is potentially a reason why brain size changes more slowly than e.g. bone size. (Also there are small structural changes in ML algorithms as you scale them up, not sure how much you want to push the analogy but they feel fairly similar.)

[Yudkowsky][15:17]

it also seems to me like the chimp->human difference is in basically the same ballpark of the effec of brain size within humans, given modest adaptations for culture

this part also seems pretty blatantly false to me
is there, like, a smooth graph that you looked at there?

[Christiano][15:18]

I think the extrapolated difference would be about 4 standard deviations, so we are comparing a chimp to an IQ 40 human

[Yudkowsky][15:18]

I'm really not sure how much of a fair comparison that is
IQ 40 humans in our society may be mostly sufficiently-damaged humans, not scaled-down humans

[Christiano][15:19]

doesn't seem easy, but the point is that the extrapolated difference is huge, it corresponds to completely debilitating developmental problems

[Yudkowsky][15:19]

if you do enough damage to a human you end up with, for example, a coma victim who's not competitive with other primates at all

[Christiano][15:19]

yes, that's more than 4 SD down
I agree with this general point
I'd guess I just have a lot more respect for chimps than you do

[Yudkowsky][15:20]

I feel like I have a bunch of respect for chimps but more respect for humans
like, that stuff humans do
that is really difficult stuff!
it is not just scaled-up chimpstuff!

[Christiano][15:21]

Carl convinced me chimps wouldn't go to space, but I still really think it's about domesticity and cultural issues rather than intelligence

[Yudkowsky][15:21]

the chimpstuff is very respectable but there is a whole big layer cake of additional respect on top

[Christiano][15:21]

not a prediction to be resolved until after the singularity
I mean, the space prediction isn't very confident 😊

and it involved a very large planet of apes

9.6. Architectural innovation in AI and in evolutionary history

[Yudkowsky][15:22]

I feel like if GPT-based systems saturate and require *any* architectural innovation rather than Stack M Layers to get much further, this is a pre-Singularity point of observation which favors humans probably being more qualitatively different from chimp-LCA

(LCA=last common ancestor)

[Christiano][15:22]

any seems like a kind of silly bar?

[Yudkowsky][15:23]

because single architectural innovations are allowed to have large effects!

[Christiano][15:23]

like there were already small changes to normalization from GPT-2 to GPT-3, so isn't it settled?

[Yudkowsky][15:23]

natural selection can't afford to deploy that many of them!

[Christiano][15:23]

and the model really eventually won't work if you increase layers but don't fix the normalization, there are severe problems that only get revealed at high scale

[Yudkowsky][15:23]

that I wouldn't call architectural innovation

transformers were

this is a place where I would not discuss specific ideas because I do not actually want this event to occur

[Christiano][15:24]

sure

have you seen a graph of LSTM scaling vs transformer scaling?

I think LSTM with ongoing normalization-style fixes lags like 3x behind transformers on language modeling

[Yudkowsky][15:25]

no, does it show convergence at high-enough scales?

[Christiano][15:25]

figure 7 here: <https://arxiv.org/pdf/2001.08361.pdf>

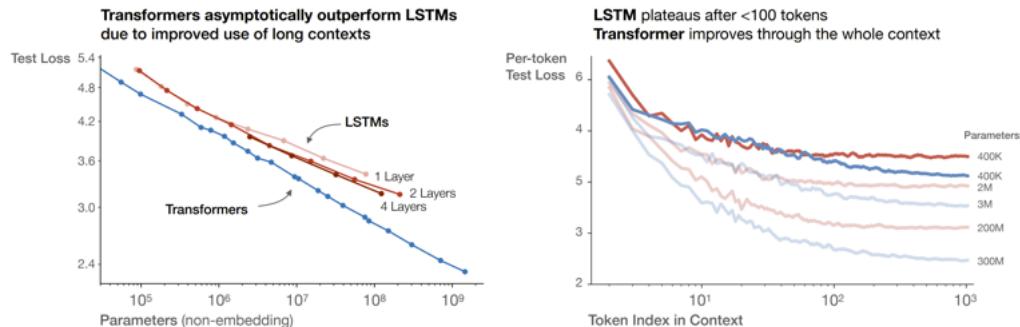


Figure 7

[Yudkowsky][15:26]

yeah... I unfortunately would rather not give other people a sense for which innovations are obviously more of the same and which innovations obviously count as qualitative

[Christiano][15:26]

I think smart money is that careful initialization and normalization on the RNN will let it keep up for longer

anyway, I'm very open to differences like LSTM vs transformer between humans and 3x-smaller-brain ancestors, as long as you are open to like 10 similar differences further back in the evolutionary history

[Yudkowsky][15:28]

what if there's 27 differences like that and 243 differences further back in history?

[Christiano][15:28]

sure

[Yudkowsky][15:28]

is that a distinctly Yudkowskian view vs a Paul view...

apparently not

I am again feeling confused about cruxes

[Christiano][15:29]

I mean, 27 differences like transformer vs LSTM isn't actually plausible, so I guess we could talk abou

[Cotra][15:30]

Here's a potential crux articulation that ties it back to the animals stuff: paul thinks that we first discover major algorithmic innovations that improve intelligence at a low level of intelligence, analogous to evolution discovering major architectural innovations with tiny birds and primates, and then there will be a long period of scaling up plus coming up with routine algorithmic tweaks to get to the high level analogous to evolution schlepping on the same shit for a long time to get to humans. analogously, he thinks when big innovations come onto the scene the actual product is crappy af (e.g. wright brother plane), and it needs a ton of work to scale up to usable and then to great.

you both seem to think both evolution and tech history consistently point in your direction

[Christiano][15:33]

that sounds vaguely right, I guess the important part of "routine" is "vaguely predictable," like you mostly work your way down the low-hanging fruit (including new fruit that becomes more important to you scale), and it becomes more and more predictable the more people are working on it and the longer you've been at it

and deep learning is already reasonably predictable (i.e. the impact of successive individual architectural changes is smaller, and law of large numbers is doing its thing) and is getting more so, I just expect that to continue

[Cotra][15:34]

yeah, like it's a view that points to using data that relates effort to algorithmic progress and using that to predict future progress (in combination with predictions of future effort)

[Christiano][15:35]

yeah

and for my part, it feels like this is how most technologies look and also how current ML progress looks

[Cotra][15:36]

and also how evolution looks, right?

[Christiano][15:37]

you aren't seeing big jumps in translation or in self-driving cars or in image recognition, you are just seeing a long slog, and you see big jumps in areas where few people work (usually up to levels that are not in fact that important, which is very correlated with few people working there)

I don't know much about evolution, but it at least looks very consistent with what I know and the fact eliezer cites

(not merely consistent, but "explains the data just about as well as the other hypotheses on offer")

9.7. Styles of thinking in forecasting

[Yudkowsky][15:38]

I do observe that this would seem, on the surface of things, to describe the entire course of natural selection up until about 20K years ago, if you were looking at surface impacts

[Christiano][15:39]

by 20k years ago I think it's basically obvious that you are tens of thousands of years from the singularity

like, I think natural selection is going crazy with the brains by millions of years ago, and by hundreds thousands of years ago humans are going crazy with the culture, and by tens of thousands of years the culture thing has accelerated and is almost at the finish line

[Yudkowsky][15:41]

really? I don't know if I would have been able to call that in advance if I'd never seen the future or an other planets. I mean, maybe, but I sure would have been extrapolating way out onto a further limb than I'm going here.

[Christiano][15:41]

Yeah, I agree singularity is way more out on a limb---or like, where the singularity stops is more uncertain since that's all that's really at issue from my perspective

but the point is that everything is clearly crazy in historical terms, in the same way that 2000 is crazy even if you don't know where it's going

and the timescale for the crazy changes is tens of thousands of years

[Yudkowsky][15:42]

I frankly model that, had I made any such prediction 20K years ago of hominids being able to pull off moon landings or global warming - never mind the Singularity - I would have faced huge pushback from many EAs, such as, for example, Robin Hanson, and you.

[Christiano][15:42]

like I think this can't go on would have applied just as well:

<https://www.lesswrong.com/posts/5FZxhd16hZp8QwK7k/this-can-t-go-on>

I don't think that's the case at all

and I think you still somehow don't understand my position?

[Yudkowsky][15:43]

<https://www.lesswrong.com/posts/XQirei3crsLxsCQoi/surprised-by-brains> is my old entry here

[Christiano][15:43]

like, what is the move I'm making here, that you think I would have made in the past?

and would have led astray?

[Yudkowsky][15:44]

I sure do feel in a deeper sense that I am trying very hard to account for perspective shifts in how unpredictable the future actually looks at the time, and the Other is looking back at the past and organizing it neatly and expecting the future to be that neat

[Christiano][15:45]

I don't even feel like I'm expecting the future to be neat

are you just saying you have a really broad distribution over takeoff speed, and that "less than a month gets a lot of probability because lots of numbers are less than a month?

[Yudkowsky][15:47]

not exactly?

[Christiano][15:47]

in what way is your view the one that is preferred by things being messy or unpredictable?

like, we're both agreeing X will eventually happen, and I'm making some concrete prediction about h some other X' will happen first, and that's the kind of specific prediction that's likely to be wrong?

[Yudkowsky][15:48]

more like, we sure can tell a story today about how normal and predictable AlphaGo was, but we can always tell stories like that about the past. I do not particularly recall the AI field standing up one year before AlphaGo and saying "It's time, we're coming for the 8-dan pros this year and we're gonna be world champions a year after that." (Which took significantly longer in chess, too, matching my other thesis about how these slides are getting steeper as we get closer to the end.)

[Christiano][15:49]

it's more like, you are offering AGZ as an example of why things are crazy, and I'm doubtful / think it's pretty lame

maybe I don't understand how it's functioning as bayesian evidence

for what over what

[Yudkowsky][15:50]

I feel like the whole smoothness-reasonable-investment view, if evaluated on Earth 5My ago *without benefit of foresight*, would have dismissed the notion of brains overtaking evolution; evaluated 1My ago it would have dismissed the notion of brains overtaking evolution; evaluated 20Ky ago, it would have barely started to acknowledge that brains were doing anything interesting at all, but pointed out how the hominids could still only eat as much food as their niche offered them and how the cute little handaxes did not begin to compare to livers and wasp stings.

there is a style of thinking that says, "wow, yeah, people in the past sure were surprised by stuff, oh, wait, I'm also in the past, aren't I, I am one of those people"

and a view where you look back from the present and think about how reasonable the past all seems now, and the future will no doubt be equally reasonable

[Christiano][15:52]

(the AGZ example may fall flat, because the arguments we are making about it now we were also making in the past)

[Yudkowsky][15:52]

I am not sure this is resolvable, but it is among my primary guesses for a deep difference in believed styles of thought

[Christiano][15:52]

I think that's a useful perspective, but still don't see how it favors your bottom line

[Yudkowsky][15:53]

where I look at the style of thinking you're using, and say, not, "well, that's invalidated by a technical error on line 3 even on Paul's own terms" but "isn't this obviously a whole style of thought that never works and ends up unrelated to reality"

I think the first AlphaGo was the larger shock, AlphaGo Zero was a noticeable but more mild shock or account of how it showed the end of game programming and not just the end of Go

[Christiano][15:54]

sorry, I lumped them together

[Yudkowsky][15:54]

it didn't feel like the same level of surprise; it was preceded by then

the actual accomplishment may have been larger in an important sense, but a lot of the - epistemic landscape of lessons learned? - is about the things that surprise you at the time

[Christiano][15:55]

also AlphaGo was also quite easy to see coming after this paper (as was discussed extensively at *the time*): <https://www.cs.toronto.edu/~cmaddis/pubs/deepgo.pdf>

[Yudkowsky][15:55]

Paul, are you on the record as arguing with me that AlphaGo will win at Go because it's predictably on trend?

back then?

[Cotra][15:55]

Hm, it sounds like Paul is saying "I do a trend extrapolation over long time horizons and if things seem to be getting faster and faster I expect they'll continue to accelerate; this extrapolation if done 100k years ago would have seen that things were getting faster and faster and projected singularity within 100s of K years"

Do you think Paul is in fact doing something other than the trend extrapolation he says he's doing, or that he would have looked at a different less informative trend than the one he says he would have looked at something else?

[Christiano][15:56]

my methodology for answering that question is looking at LW comments mentioning go by me, can see if it finds any

[Yudkowsky][15:56]

Different less informative trend, is most of my suspicion there?

though, actually, I should revise that, I feel like relatively little of the WHA was AlphaGo v2 whose name forget beating Lee Se-dol, and most was in the revelation that v1 beat the high-dan pro whose name forget.

Paul having himself predicted anything at all like this would be the actually impressive feat

that would cause me to believe that the AI world is more regular and predictable than I experienced it as, if you are paying more attention to ICLR papers than I do

9.8. Moravec's prediction

[Cotra][15:58]

And jtbc, the trend extrap paul is currently doing is something like:

- Look at how effort leads to hardware progress measured in FLOP/\$ and software progress measured in stuff like "FLOP to do task X" or "performance on benchmark Y"
- Look at how effort in the ML industry as a whole is increasing, project forward with maybe some adjustments for thinking markets are more inefficient now and will be less inefficient later

and this is the wrong trend, because he shouldn't be looking at hardware/software progress across the whole big industry and should be more open to an upset innovation coming from an area with a small number of people working on it?

and he would have similarly used the wrong trends while trying to do trend extrap in the past?

[Yudkowsky][15:59]

because I feel like this general style of thought doesn't work when you use it on Earth generally, and then fails extremely hard if you try to use it on Earth before humans to figure out where the hominids are going because that phenomenon is Different from Previous Stuff

like, to be clear, I have seen this used well on solar

I feel like I saw some people calling the big solar shift based on graphs, before that happened

I have seen this used great by Moravec on computer chips to predict where computer chips would be 2012

and also witnessed Moravec *completely failing* as soon as he tried to derive *literally anything but the graph itself* namely his corresponding prediction for human-equivalent AI in 2012 (I think, maybe it was 2010) or something

[Christiano][16:02]

(I think in his 1988 book Moravec estimated human-level AI in ~2030, not sure if you are referring to some earlier prediction?)

[Yudkowsky][16:02]

(I have seen Ray Kurzweil project out Moore's Law to the \$1,000,000 human brain in, what was it, 20 followed by the \$1000 human brain in 2035 and the \$1 human brain in 2045, and when I asked Ray whether machine superintelligence might shift the graph at all, he replied that machine superintelligence was precisely how the graph would be able to continue on trend. This indeed is silly than EAs.)

[Cotra][16:03]

moravec's prediction appears to actually be around 2025, looking at his hokey graph?
<https://jetpress.org/volume1/moravec.htm>



[Yudkowsky][16:03]

but even there, it does feel to me like there is a commonality between Kurzweil's sheer graph-worship and difficulty in appreciating the graphs as surface phenomena that are less stable than deep phenomena, and something that Hanson was doing wrong in the foom debate

[Cotra][16:03]

which is...like, your timelines?

[Yudkowsky][16:04]

that's 1998

Mind Children in 1988 I am pretty sure had an earlier prediction

[Christiano][16:04]

I should think you'd be happy to bet against me on basically any prediction, shouldn't you?

[Yudkowsky][16:05]

any prediction that sounds narrow and isn't like "this graph will be on trend in 3 more years"

...maybe I'm wrong, an online source says Mind Children in 1988 predicted AGI in "40 years" but I sur do seem to recall an extrapolated graph that reached "human-level hardware" in 2012 based on an extensive discussion about computing power to duplicate the work of the retina

[Christiano][16:08]

don't think it matters too much other than for Moravec's honor, doesn't really make a big difference to the empirical success of the methodology

I think it's on page 68 if you have the physical book

[Yudkowsky][16:09]

p60 via Google Books says 10 teraops for a human-equivalent mind

[Christiano][16:09]

I have a general read of history where trend extrapolation works extraordinarily well relative to other kinds of forecasting, to the extent that the best first-pass heuristic for whether a prediction is likely to

accurate is whether it's a trend extrapolation and how far in the future it is

[Yudkowsky][16:09]

which, incidentally, strikes me as entirely plausible if you had algorithms as sophisticated as the human brain

my sense is that Moravec nailed the smooth graph of computing power going on being smooth, but that all of his predictions about the actual future were completely invalid on account of a curve interacting with his curve that he didn't know things about and so simply omitted as a step in his calculations, namely, AGI algorithms

[Christiano][16:12]

though again, from your perspective 2030 is still a reasonable bottom-line forecast that makes him one of the most accurate people at that time?

[Yudkowsky][16:12]

you could be right about all the local behaviors that your history is already shouting out at you as having smooth curve (where by "local" I do mean to exclude stuff like world GDP extrapolated into the indefinite future) and the curves that history isn't shouting at you will tear you down

[Christiano][16:12]

(I don't know if he even forecast that)

[Yudkowsky][16:12]

I don't remember that part from the 1988 book

my memory of the 1988 book is "10 teraops, based on what it takes to rival the retina" and he drew a graph of Moore's Law

[Christiano][16:13]

yeah, I think that's what he did

(and got 2030)

[Yudkowsky][16:14]

"If this rate of improvement were to continue into the next century, the 10 teraops required for a humanlike computer would be available in a \$10 million supercomputer before 2010 and in a \$1,000 personal computer by 2030."

[Christiano][16:14]

or like, he says "human equivalent in 40 years" and predicts that in 50 years we will have robots with superhuman reasoning ability, not clear he's ruling out human-equivalent AGI before 40 years but I think the tone is clear

[Yudkowsky][16:15]

so 2030 for AGI on a personal computer and 2010 for AGI on a supercomputer, and I expect that on my first reading I simply discarded the former prediction as foolish extrapolation past the model collapse

had just predicted in 2010.

(p68 in "Powering Up")

[Christiano][16:15]

yeah, that makes sense

I do think the PC number seems irrelevant

[Cotra][16:16]

I think both in that book and in the 98 article he wants you to pay attention to the "very cheap human-size computers" threshold, not the "supercomputer" threshold, I think intentionally as a way to handwave in "we need people to be able to play around with these things"

(which people criticized him at the time for not more explicitly modeling iirc)

[Yudkowsky][16:17]

but! I mean! there are so many little places where the media has a little cognitive hiccup about that; decides in 1998 that it's fine to describe that retrospectively as "you predicted in 1988 that we'd have true AI in 40 years" and then the future looks less surprising than people at the time using Trend Lines were actually surprised by it!

all these little ambiguities and places where, oh, you decide retroactively that it would have made sense to look at *this* Trend Line and use it *that* way, but if you look at what people said at the time, they did actually say that!

[Christiano][16:19]

I mean, in fairness reading the book it just doesn't seem like he is predicting human-level AI in 2010 rather than 2040, but I do agree that it seems like the basic methodology (why care about the small computer thing?) doesn't really make that much sense a priori and only leads to something sane if it cancels out with a weird view

9.9. Prediction disagreements and bets

[Christiano][16:19]

anyway, I'm pretty unpersuaded by the kind of track record appeal you are making here

[Yudkowsky][16:20]

if the future goes the way I predict and yet anybody somehow survives, perhaps somebody will draw hyperbolic trendline on some particular chart where the trendline is retroactively fitted to events including those that occurred in only the last 3 years, and say with a great sage nod, ah, yes, that was all according to trend, nor did anything depart from trend

trend lines permit anything

[Christiano][16:20]

like from my perspective the fundamental question is whether I would do better or worse by following the kind of reasoning you'd advocate, and it just looks to me like I'd do worse, and I'd love to make a predictions about anything to help make that more clear and hindsight-proof in advance

[Yudkowsky][16:20]

you just look into the past and find a line you can draw that ended up where reality went

[Christiano][16:21]

it feels to me like you really just waffle on almost any prediction about the before-end-of-days

[Yudkowsky][16:21]

I don't think I know a lot about the before-end-of-days

[Christiano][16:21]

like if you make a prediction I'm happy to trade into it, or you can pick a topic and I can make a prediction and you can trade into mine

[Cotra][16:21]

but you know enough to have strong timing predictions, e.g. your bet with caplan

[Yudkowsky][16:21]

it's daring enough that I claim to know anything about the Future at all!

[Cotra][16:21]

surely with that difference of timelines there should be some pre-2030 difference as well

[Christiano][16:21]

but you are the one making the track record argument against my way of reasoning about things!

how does that not correspond to believing that your predictions are better!

what does that mean?

[Yudkowsky][16:22]

yes and if you say something narrow enough or something that my model does at least vaguely push against, we should bet

[Christiano][16:22]

my point is that I'm willing to make a prediction about any old thing, you can name your topic

I think the way I'm reasoning about the future is just better in general

and I'm going to beat you on whatever thing you want to bet on

[Yudkowsky][16:22]

but if you say, "well, Moore's Law on trend, next 3 years", then I'm like, "well, yeah, sure, since I don't feel like I know anything special about that, that would be my prediction too"

[Christiano][16:22]

sure

you can pick the topic

pick a quantity

or a yes/no question

or whatever

[Yudkowsky][16:23]

you may know better than I would where your Way of Thought makes strong, narrow, or unusual predictions

[Christiano][16:23]

I'm going to trend extrapolation everywhere

spoiler

[Yudkowsky][16:23]

okay but any superforecaster could do that and I could do the same by asking a superforecaster

[Cotra][16:24]

but there must be places where you'd strongly disagree w the superforecaster

since you disagree with them eventually, e.g. >2/3 doom by 2030

[Bensinger][18:40] (Nov. 25 follow-up comment)

">2/3 doom by 2030" isn't an actual Eliezer-prediction, and is based on a misunderstanding of something Eliezer said. See [Eliezer's comment on LessWrong](#).

[Yudkowsky][16:24]

in the terminal phase, sure

[Cotra][16:24]

right, but there are no disagreements before jan 1 2030?

no places where you'd strongly defy the superforecasters/trend extrap?

[Yudkowsky][16:24]

superforecasters were claiming that AlphaGo had a 20% chance of beating Lee Se-dol and I didn't disagree with that at the time, though as the final days approached I became nervous and suggested

a friend that they buy out of a bet about that

[Cotra][16:25]

what about like whether we get some kind of AI ability (e.g. coding better than X) before end days

[Yudkowsky][16:25]

though that was more because of having started to feel incompetent and like I couldn't trust the superforecasters to know more, than because I had switched to a confident statement that AlphaGo would win

[Cotra][16:25]

seems like EY's deep intelligence / insight-oriented view should say something about what's not poss before we get the "click" and the FOOM

[Christiano][16:25]

I mean, I'm OK with either (i) evaluating arguments rather than dismissive and IMO totally unjustified track record, (ii) making bets about stuff

I don't see how we can both be dismissing things for track record reasons and also not disagreeing about things

if our methodologies agree about all questions before end of days (which seems crazy to me) then surely there is no track record distinction between them...

[Cotra: ]

[Cotra][16:26]

do you think coding models will be able to 2x programmer productivity before end days? 4x?

what about hardware/software R&D wages? will they get up to \$20m/yr for good ppl?

will someone train a 10T param model before end days?

[Christiano][16:27]

things I'm happy to bet about: economic value of LMs or coding models at 2, 5, 10 years, benchmark performance of either, robotics, wages in various industries, sizes of various industries, compute/\$, someone else's views about "how ML is going" in 5 years

maybe the "any GDP acceleration before end of days?" works, but I didn't like how you don't win until the end of days

[Yudkowsky][16:28]

okay, so here's an example place of a *weak* general Yudkowskian prediction, that is weaker than terminal-phase stuff of the End Days: (1) I predict that cycles of 'just started to be able to do Narrow Thing -> blew past upper end of human ability at Narrow Thing' will continue to get shorter, the same way that, I think, this happened faster with Go than with chess.

[Christiano][16:28]

great, I'm totally into it

what's a domain?

coding?

[Yudkowsky][16:28]

Does Paul disagree? Can Paul point to anything equally specific out of Paul's viewpoint?

[Christiano][16:28]

benchmarks for LMs?

robotics?

[Yudkowsky][16:28]

well, for these purposes, we do need some Elo-like ability to measure at all where things are relative humans

[Cotra][16:29]

problem-solving benchmarks for code?

MATH benchmark?

[Christiano][16:29]

well, for coding and LM'ing we have lots of benchmarks we can use

[Yudkowsky][16:29]

this unfortunately does feel a bit different to me from Chess benchmarks where the AI is playing the whole game; Codex is playing part of the game

[Christiano][16:29]

in general the way I'd measure is by talking about how fast you go from "weak human" to "strong human" (e.g. going from top-10,000 in chess to top-10 or whatever, going from jobs doable by \$50k/year engineer to \$500k/year engineer...)

[Yudkowsky][16:30]

golly, that sounds like a viewpoint very favorable to mine

[Christiano][16:30]

what do you mean?

that way of measuring would be favorable to your viewpoint?

[Yudkowsky][16:31]

if we measure how far it takes AI to go past different levels of paying professionals, I expect that the Chess duration is longer than the Go duration and that by the time Codex is replacing a most paid

\$50k/year programmers the time to replacing a most programmers paid as much as a top Go player
be pretty darned short

[Christiano][16:31]

top Go players don't get paid, do they?

[Yudkowsky][16:31]

they tutor students and win titles

[Christiano][16:31]

but I mean, they are like low-paid engineers

[Yudkowsky][16:31]

yeah that's part of the issue here

[Christiano][16:31]

I'm using wages as a way to talk about the distribution of human abilities, not the fundamental number

[Yudkowsky][16:32]

I would expect something similar to hold over going from low-paying welder to high-paying welder

[Christiano][16:32]

like, how long to move from "OK human" to "pretty good human" to "best human"

[Cotra][16:32]

says salary of \$350k/yr for lee: <https://www.fameranker.com/lee-sedol-net-worth>

[Yudkowsky][16:32]

but I also mostly expect that AIs will not be allowed to weld things on Earth

[Cotra][16:32]

why don't we just do an in vitro benchmark instead of wages?

[Christiano][16:32]

what, machines already do virtually all welding?

[Cotra][16:32]

just pick a benchmark?

[Yudkowsky][16:33]

yoouuuu do not want to believe sites like that (fameranker)

[Christiano][16:33]

yeah, I'm happy with any benchmark, and then we can measure various human levels at that benchmark

[Cotra][16:33]

what about MATH? <https://arxiv.org/abs/2103.03874>

[Christiano][16:34]

also I don't know what "shorter and shorter" means, the time in go and chess was decades to move from "strong amateur" to "best human," I do think these things will most likely be shorter than decades seems like we can just predict concrete #s though

[Cotra: ]

like I can say how long I think it will take to get from "median high schooler" to "IMO medalist" and you can bet against me?

and if we just agree about all of those predictions then again I'm back to being very skeptical of a claimed track record difference between our models

(I do think that it's going to take years rather than decades on all of these things)

[Yudkowsky][16:36]

possibly! I worry this ends up in a case where Katja or Luke or somebody goes back and collects data about "amateur to pro performance times" and Eliezer says "Ah yes, these are shortening over time, as I predicted" and Paul is like "oh, well, I predict they continue to shorten on this trend drawn from the data" and Eliezer is like "I guess that could happen for the next 5 years, sure, sounds like something superforecaster would predict as default"

[Cotra][16:37]

i'm pretty sure paul's methodology here will just be to look at the MATH perf trend based on model size and combine with expectations of when ppl will make big enough models, not some meta trend thing like that?

[Yudkowsky][16:37]

so I feel like... a bunch of what I feel is the real disagreement in our models, is a bunch of messy stuff. Suddenly Popping Up one day and then Eliezer is like "gosh, I sure didn't predict that" and Paul is like "somebody could have totally predicted that" and Eliezer is like "people would say exactly the same thing after the world ended in 3 minutes"

if we've already got 2 years of trend on a dataset, I'm not necessarily going to predict the trend breaking

[Cotra][16:38]

hm, you're presenting your view as more uncertain and open to anything here than paul's view, but in fact it's picking out a narrower distribution. you're more confident in powerful AGI soon

[Christiano][16:38]

seems hard to play the "who is more confident?" game

[Cotra][16:38]

so there should be some places where you make a strong positive prediction paul disagrees with

[Yudkowsky][16:39]

I might want to buy options on a portfolio of trends like that, if Paul is willing to sell me insurance against all of the trends breaking upward at a lower price than I think is reasonable

I mean, from my perspective Paul is the one who seems to think the world is well-organized and predictable in certain ways

[Christiano][16:39]

yeah, and you are saying that I'm overconfident about that

[Yudkowsky][16:39]

I keep wanting Paul to go on and make narrower predictions than I do in that case

[Christiano][16:39]

so you should be happy to bet with me about *anything*

and I'm letting you pick anything at all you want to bet about

[Cotra][16:40]

i mean we could do a portfolio of trends like MATH and you could bet on at least a few of them having strong surprises in the sooner direction

but that means we could just bet about MATH and it'd just be higher variance

[Yudkowsky][16:40]

ok but you're not going to sell me cheap options on sharp declines in the S&P 500 even though in a very reasonable world there would not be any sharp declines like that

[Christiano][16:41]

if we're betting \$ rather than bayes points, then yes I'm going to weigh worlds based on the value of in those worlds

[Cotra][16:41]

wouldn't paul just sell you options at the price the options actually trade for? i don't get it

[Christiano][16:41]

but my sense is that I'm just generally across the board going to be more right than you are, and I'm frustrated that you just keep saying that "people like me" are wrong about stuff

[Yudkowsky][16:41]

Paul's like "we'll see smooth behavior in the end days" and I feel like I should be able to say "then Paul sell me cheap options against smooth behavior now" but Paul is just gonna wanna sell at market price

[Christiano][16:41]

and so I want to hold you to that by betting about anything

ideally just tons of stuff

random things about what AI will be like, and other technologies, and regulatory changes

[Cotra][16:42]

paul's view doesn't seem to imply that he should value those options less than the market

he's more EMH-y than you not less

[Yudkowsky][16:42]

but then the future should *behave like that market*

[Christiano][16:42]

what do you mean?

[Yudkowsky][16:42]

it should have options on wild behavior that are not cheap!

[Christiano][16:42]

you mean because people want \$ more in worlds where the market drops a lot?

I don't understand the analogy

[Yudkowsky][16:43]

no, because jumpy stuff happens more than it would in a world of ideal agents

[Cotra][16:43]

I think EY is saying the non-cheap option prices are because P(sharp declines) is pretty high

[Christiano][16:43]

ok, we know how often markets jump, if that's the point of your argument can we just talk about that directly?

[Yudkowsky][16:43]

or sharp rises, for that matter

[Christiano][16:43]

(much lower than option prices obviously)

I'm probably happy to sell you options for sharp rises

I'll give you better than market odds in that direction

that's how this works

[Yudkowsky][16:44]

now I am again confused, for I thought you were the one who expected world GDP to double in 4 years at some point

and indeed, drew such graphs with the rise suggestively happening earlier than the sharp spike

[Christiano][16:44]

yeah, and I have exposure to that by buying stocks, options prices are just a terrible way of tracking these things

[Yudkowsky][16:44]

suggesting that such a viewpoint is generally favorable to near timelines for that

[Christiano][16:44]

I mean, I have bet a *lot* of money on AI companies doing well

well, not compared to the EA crowd, but compared to my meager net worth 😊

and indeed, it has been true so far

and I'm continuing to make the bet

it seems like on your view it should be surprising that AI companies just keep going up

aren't you predicting them not to get to tens of trillions of valuation before the end of days?

[Yudkowsky][16:45]

I believe that Nate, of a generally Yudkowskian view, did the same (bought AI companies). and I focused my thoughts elsewhere, because somebody needs to, but did happen to buy my first S&P 500 on its exact minimum in 2020

[Christiano][16:46]

point is, that's how you get exposure to the crazy growth stuff with continuous ramp-ups

and I'm happy to make the bet on the market

or on other claims

I don't know if my general vibe makes sense here, and why it seems reasonable to me that I'm just happy to bet on anything

as a way of trying to defend my overall attack

and that if my overall epistemic approach is vulnerable to some track record objection, then it seems like it ought to be possible to win here

9.10. Prediction disagreements and bets: Standard superforecaster techniques

[Cotra][16:47]

I'm still kind of surprised that Eliezer isn't willing to bet that there will be a faster-than-Paul expects trend break on MATH or whatever other benchmark. Is it just the variance of MATH being one benchmark? Would you make the bet if it were 6?

[Yudkowsky][16:47]

a large problem here is that both of us tend to default strongly to superforecaster standard technique

[Christiano][16:47]

it's true, though it's less true for longer things

[Cotra][16:47]

but you think the superforecasters would suck at predicting end days because of the surface trends thing!

[Yudkowsky][16:47]

before I bet against Paul on MATH I would want to know that Paul wasn't arriving at the same default use, which might be drawn from trend lines there, or from a trend line in trend lines

I mean the superforecasters did already suck once in my observation, which was AlphaGo, but I did not bet against them there, I bet with them and then updated afterwards

[Christiano][16:48]

I'd mostly try to eyeball how fast performance was improving with size; I'd think about difficulty effect (where e.g. hard problems will be flat for a while and then go up later, so you want to measure performance on a spectrum of difficulties)

[Cotra][16:48]

what if you bet against a methodology instead of against paul's view? the methodology being the one described above, of looking at the perf based on model size and then projecting model size increases cost?

[Christiano][16:48]

seems safer to bet against my view

[Cotra][16:48]

yeah

[Christiano][16:48]

mostly I'd just be eyeballing size, thinking about how much people will in fact scale up (which would great to factor out if possible), assuming performance trends hold up

are there any other examples of surface trends vs predictable deep changes, or is AGI the only one?
(that you have thought a lot about)

[Cotra][16:49]

yeah seems even better to bet on the underlying "will the model size to perf trends hold up or break upward"

[Yudkowsky][16:49]

so from my perspective, there's this whole thing where *unpredictably* something breaks above trend because the first way it got done was a way where somebody could do it faster than you expected

[Christiano][16:49]

(makes sense for it to be the domain where you've thought a lot)

you mean, it's unpredictable what will break above trend?

[Cotra][16:49]

[IEM](#) has a financial example

[Yudkowsky][16:49]

I mean that I could not have said "Go will break above trend" in 2015

[Christiano][16:49]

yeah

ok, here's another example

[Yudkowsky][16:50]

it feels like if I want to make a bet with imaginary Paul in 2015 then I have to bet on a portfolio
and I also feel like as soon as we make it that concrete, Paul does not want to offer me things that I w
to bet on

because Paul is also like, sure, something might break upward

I remark that I have for a long time been saying that I wish Paul had more concrete images and
examples attached to *a lot of his stuff*

[Cotra][16:51]

surely the view is about the probability of each thing breaking upward. or the expected number from basket

[Christiano][16:51]

I mean, if you give me any way of quantifying how much stuff breaks upwards we have a bet

[Cotra][16:51]

not literally that one single thing breaks upward

[Christiano][16:51]

I don't understand how concreteness is an accusation here, I've offered 10 quantities I'd be happy to about, and also allowed you to name literally any other quantity you want

and I agree that we mostly agree about things

[Yudkowsky][16:52]

and some of my sense here is that if Paul offered a portfolio bet of this kind, I might not take it myself but EAs who were better at noticing their own surprise might say, "Wait, that's how unpredictable Paul thinks the world is?"

so from my perspective, it is hard to know specific anti-superforecaster predictions that happen long before terminal phase, and I am not sure we are really going to get very far there.

[Christiano][16:53]

but you agree that the eventual prediction is anti-superforecaster?

[Yudkowsky][16:53]

both of us probably have quite high inhibitions against selling conventionally priced options that are not what a superforecaster would price them as

[Cotra][16:53]

why does it become so much easier to know these things and go anti-superforecaster at terminal phase?

[Christiano][16:53]

I assume you think that the superforecasters will continue to predict that big impactful AI applications are made by large firms spending a lot of money, even through the end of days

I do think it's very often easy to beat superforecasters in-domain

like I expect to personally beat them at most ML prediction

and so am also happy to do bets where you defer to superforecasters on arbitrary questions and I be against you

[Yudkowsky][16:54]

well, they're anti-prediction-market in the sense that, at the very end, bets can no longer settle. I've been surprised of late by how much AGI ruin seems to be sneaking into common knowledge; perhaps the terminal phase the superforecasters will be like, "yep, we're dead". I can't even say that in this case Paul will disagree with them, because I expect the state on alignment to be so absolutely awful that even Paul is like "You were not supposed to do it that way" in a very sad voice.

[Christiano][16:55]

I'm just thinking about takeoff speeds here

I do think it's fairly likely I'm going to be like "oh no this is bad" (maybe 50%?), but not that I'm going to expect fast takeoff

and similarly for the superforecasters

9.11. Prediction disagreements and bets: Late-stage predictions, and betting against superforecasters

[Yudkowsky][16:55]

so, one specific prediction you made, sadly close to terminal phase but not much of a surprise there, that the world economy must double in 4 years before the End Times are permitted to begin

[Christiano][16:56]

well, before it doubles in 1 year...

I think most people would call the 4 year doubling the end times

[Yudkowsky][16:56]

this seems like you should also be able to point to some least impressive thing that is not permitted to occur before WGDP has doubled in 4 years

[Christiano][16:56]

and it means that the normal planning horizon includes the singularity

[Yudkowsky][16:56]

it may not be much but we would be *moving back* the date of first concrete disagreement

[Christiano][16:57]

I can list things I don't think would happen first, since that's a ton

[Yudkowsky][16:57]

and EAs might have a little bit of time in which to say "Paul was falsified, uh oh"

[Christiano][16:57]

the only things that aren't permitted are the ones that would have caused the world economy to dou in 4 years

[Yudkowsky][16:58]

and by the same token, there are things Eliezer thinks you are probably not going to be able to do before you slide over the edge. a portfolio of these will have some losing options because of adverse selection against my errors of what is hard, but if I lose more than half the portfolio, this may said to a bad sign for Eliezer.

[Christiano][16:58]

(though those can happen at the beginning of the 4 year doubling)

[Yudkowsky][16:58]

this is unfortunately *late* for falsifying our theories but it would be *progress* on a kind of bet against e other

[Christiano][16:59]

but I feel like the things I'll say are like fully automated construction of fully automated factories at 1-year turnarounds, and you're going to be like "well duh"

[Yudkowsky][16:59]

...unfortunately yes

[Christiano][16:59]

the reason I like betting about numbers is that we'll probably just disagree on any given number

[Yudkowsky][16:59]

I don't think I *know* numbers.

[Christiano][16:59]

it does seem like a drawback that this can just turn up object-level differences in knowledge-of-numb more than deep methodological advantages

[Yudkowsky][17:00]

the last important number I had a vague suspicion I might know was that Ethereum ought to have a significantly larger market cap in pre-Singularity equilibrium.

and I'm not as sure of that one since El Salvador supposedly managed to use Bitcoin L2 Lightning.

(though I did not fail to act on the former belief)

[Christiano][17:01]

do you see why I find it weird that you think there is this deep end-times truth about AGI, that is very different from a surface-level abstraction and that will take people like Paul by surprise, without think there are other facts like that about the world?

I do see how this annoying situation can come about
and I also understand the symmetry of the situation

[Yudkowsky][17:02]

we unfortunately both have the belief that the present world looks a lot like our being right, and therefore that the other person ought to be willing to bet against default superforecasterish projectio

[Cotra][17:02]

paul says that *he* would bet against superforecasters too though

[Christiano][17:02]

I would in ML

[Yudkowsky][17:02]

like, where specifically?

[Christiano][17:02]

or on any other topic where I can talk with EAs who know about the domain in question

I don't know if they have standing forecasts on things, but e.g.: (i) benchmark performance, (ii) indus size in the future, (iii) how large an LM people will train, (iv) economic impact of any given ML system like codex, (v) when robotics tasks will be plausible

[Yudkowsky][17:03]

I have decided that, as much as it might gain me prestige, I don't think it's actually the right thing for me to go spend a bunch of character points on the skills to defeat superforecasters in specific domains and then go around doing that to prove my epistemic virtue.

[Christiano][17:03]

that seems fair

[Yudkowsky][17:03]

you don't need to bet with *me* to prove your epistemic virtue in this way, though

okay, but, if I'm allowed to go around asking Carl Shulman who to ask in order to get the economic impact of Codex, maybe I can also defeat superforecasters.

[Christiano][17:04]

I think the deeper disagreement is that (i) I feel like my end-of-days prediction is also basically just a default superforecaster prediction (and if you think yours is too then we can bet about what some superforecasters will say on it), (ii) I think you are leveling a much stronger "people like paul get taken by surprise by reality" claim whereas I'm just saying that I don't like your arguments

[Yudkowsky][17:04]

it seems to me like the contest should be more like our intuitions in advance of doing that

[Christiano][17:04]

yeah, I think that's fine, and also cheaper since research takes so much time
I feel like those asymmetries are pretty strong though

9.12. Self-duplicating factories, AI spending, and Turing test variants

[Yudkowsky][17:05]

so, here's an idea that is less epistemically virtuous than our making Nicely Resolvable Bets
what if we, like, talked a bunch about our off-the-cuff senses of where various AI things are going in t
next 3 years
and then 3 years later, somebody actually reviewed that

[Christiano][17:06]

I do think just saying a bunch of stuff about what we expect will happen so that we can look back on
would have a significant amount of the value

[Yudkowsky][17:06]

and any time the other person put a thumbs-up on the other's prediction, that prediction coming true
was not taken to distinguish them

[Cotra][17:06]

i'd suggest doing this in a format other than discord for posterity

[Yudkowsky][17:06]

even if the originator was like HOW IS THAT ALSO A PREDICTION OF YOUR THEORY
well, Discord has worked better than some formats

[Cotra][17:07]

something like a spreadsheet seems easier for people to look back on and score and stuff
discord transcripts are pretty annoying to read

[Yudkowsky][17:08]

something like a spreadsheet seems liable to be high-cost and not actually happen

[Christiano][17:08]

I think a conversation is probably easier and about as good for our purposes though?

[Cotra][17:08]

ok fair

[Yudkowsky][17:08]

I think money can be inserted into humans in order to turn Discord into spreadsheets

[Christiano][17:08]

and it's possible we will both think we are right in retrospect

and that will also be revealing

[Yudkowsky][17:09]

but, besides that, I do want to boop on the point that I feel like Paul should be able to predict intuitively rather than with necessity, things that should not happen before the world economy doubled in 4 years

[Christiano][17:09]

it may also turn up some quantitative differences of view

there are lots of things I think won't happen before the world economy has doubled in 4 years

[Yudkowsky][17:09]

because on my model, as we approach the end times, AI was still pretty partial and also the world economy was lolling most of the inputs a sensible person would accept from it and prototypes weren't being commercialized and stuff was generally slow and messy

[Christiano][17:09]

prototypes of factories building factories in <2 years

[Yudkowsky][17:10]

"AI was still pretty partial" leads it to not do interesting stuff that Paul can rule out

[Christiano][17:10]

like I guess I think Tesla will try, and I doubt it will be just Tesla

[Yudkowsky][17:10]

but the other parts of that permit AI to do interesting stuff that Paul can rule out

[Christiano][17:10]

automated researchers who can do ML experiments from 2020 without human input

[Yudkowsky][17:10]

okay, see, that whole "factories building factories" thing just seems so very much *after* the End Time
me

[Christiano][17:10]

yeah, we should probably only talk about cognitive work
since you think physical work will be very slow

[Yudkowsky][17:11]

okay but not just that, it's a falsifiable prediction
it is something that lets Eliezer be wrong in advance of the End Times

[Christiano][17:11]

what's a falsifiable prediction?

[Yudkowsky][17:11]

if we're in a world where Tesla is excitingly gearing up to build a fully self-duplicating factory including
its mining inputs and chips and solar panels and so on, we're clearly in the Paulverse and not in the
Eliezerverse!

[Christiano][17:12]

yeah
I do think we'll see that before the end times
just not before 4 year doublings

[Yudkowsky][17:12]

this unfortunately only allows you to be right, and not for me to be right, but I think there are also thi
you legit only see in the Eliezerverse!

[Christiano][17:12]

I mean, I don't think they will be doing mining for a long time because it's cheap

[Yudkowsky][17:12]

they are unfortunately late in the game but they exist at all!
and being able to state them is progress on this project!

[Christiano][17:13]

but fully-automated factories first, and then significant automation of the factory-building process
I do expect to see
I'm generally pretty bullish on industrial robotics relative to you I think, even before the crazy stuff?
but you might not have a firm view
like I expect to have tons of robots doing all kinds of stuff, maybe cutting human work in manufacturing 2x, with very modest increases in GDP resulting from that in particular

[Yudkowsky][17:13]

so, like, it doesn't surprise me very much if Tesla manages to fully automate a factory that takes in some relatively processed inputs including refined metals and computer chips, and outputs a car? and by the same token I expect that has very little impact on GDP.

[Christiano][17:14]

refined metals are almost none of the cost of the factory
and also tesla isn't going to be that vertically integrated
the fabs will separately continue to be more and more automated
I expect to have robot cars driving everywhere, and robot trucks
another 2x fall in humans required for warehouses
elimination of most brokers involved in negotiating shipping

[Yudkowsky][17:15]

if despite the fabs being more and more automated, somehow things are managing not to cost less and less, and that sector of the economy is not really growing very much, is that more like the Eliezerverse than the Paulverse?

[Christiano][17:15]

most work in finance and loan origination

[Yudkowsky][17:15]

though this is something of a peripheral prediction to AGI core issues

[Christiano][17:16]

yeah, I think if you cut the humans to do X by 2, but then the cost falls much less than the number you'd naively expect (from saving on the human labor and paying for the extra capital), then that's surprising to me

I mean if it falls half as much as you'd expect on paper I'm like "that's a bit surprising" rather than having my mind blown, if it doesn't fall I'm more surprised

but that was mostly physical economy stuff

oh wait, I was making positive predictions now, physical stuff is good for that I think?
since you don't expect it to happen?

[Yudkowsky][17:17]

...this is not your fault but I wish you'd asked me to produce my "percentage of fall vs. paper calculation" estimate before you produced yours

my mind is very whiffy about these things and I am not actually unable to deanchor on your estimate

[Christiano][17:17]

makes sense, I wonder if I should just spoiler
one benefit of discord

[Yudkowsky][17:18]

yeah that works too!

[Christiano][17:18]

a problem for prediction is that I share some background view about insane
inefficiency/inadequacy/decadence/silliness
so these predictions are all tampered by that
but still seem like there are big residual disagreements

[Yudkowsky][17:19]

sighgreat

[Christiano][17:19]

since you have way more of that than I do

[Yudkowsky][17:19]

not your fault but

[Christiano][17:19]

I think that the AGI stuff is going to be a gigantic megaproject despite that

[Yudkowsky][17:19]

I am not shocked by the AGI stuff being a gigantic megaproject
it's not above the bar of survival but, given other social optimism, it permits death with more dignity
than by other routes

[Christiano][17:20]

what if spending is this big:

Google invests \$100B training a model, total spending across all of industry is way bigger

[Yudkowsky][17:20]

ooooh

I do start to be surprised if, come the end of the world, AGI is having more invested in it than a TSMC though, not... *super* surprised?

also I am at least a little surprised before then

actually I should probably have been spoiling those statements myself but my expectation is that Paul's secret spoiler is about

\$10 trillion dollars or something equally totally shocking to an Eliezer

[Christiano][17:22]

my view on that level of spending is

it's an only slightly high-end estimate for spending by someone on a single model, but that in practice there will be ways of dividing more across different firms, and that the ontology of single-model will likely be slightly messed up (e.g. by OpenAI Five-style surgery). Also if it's that much then it likely involves big institutional changes and isn't at Google.

I read your spoiler

my estimate for total spending for the whole project of making TAI, including hardware and software manufacturing and R&D, the big datacenters, etc.

is in the ballpark of \$10T, though it's possible that it will be undercounted several times due to wage stickiness for high-end labor

[Yudkowsky][17:24]

I think that as

spending on particular AGI megaprojects starts to go past \$50 billion, it's not especially ruled out per se by things that I think I know for sure, but I feel like a third-party observer should justly start to weakly think, 'okay, this is looking at least a little like the Paulverse rather than the Eliezerverse', and as we get closer to \$10 trillion, that is not absolutely ruled out by the Eliezerverse but it was a whole lot more strongly predicted by the Paulverse, maybe something like 20x unless I'm overestimating how strongly Paul predicts that

[Christiano][17:24]

Proposed modification to the "speculate about the future to generate kind-of-predictions" methodology: we make shit up, then later revise based on points others made, and maybe also get Carl to sanity-check and decide which of his objections we agree with. Then we can separate out the "how good are our intuitions" claim (with fast feedback) from the all-things-considered how good was the "prediction"

[Yudkowsky][17:25]

okay that hopefully allows me to read Paul's spoilers... no I'm being silly. @ajeya please read all the spoilers and say if it's time for me to read his

[Cotra][17:25]

you can read his latest

[Christiano][17:25]

I'd guess it's fine to read all of them?

[Cotra][17:26]

yeah sorry that's what i meant

[Yudkowsky][17:26]

what should I say more about before reading earlier ones?

ah k

[Christiano][17:26]

My \$10T estimate was after reading yours (didn't offer an estimate on that quantity beforehand), tho that's the kind of ballpark I often think about, maybe we should just spoiler only numbers so that context is clear 😊

I think fast takeoff gets significantly more likely as you push that number down

[Yudkowsky][17:27]

so, may I now ask what starts to look to you like "oh damn I am in the Eliezerverse"?

[Christiano][17:28]

big mismatches between that AI looks technically able to do and what AI is able to do, though that's going to need a lot of work to operationalize

I think low growth of AI overall feels like significant evidence for Eliezerverse (even if you wouldn't m that prediction), since I'm forecasting it rising to absurd levels quite fast whereas your model is consistent with it staying small

some intuition about AI looking very smart but not able to do much useful until it has the whole pictu I guess this can be combined with the first point to be something like---AI looks really smart but it's ju not adding much value

all of those seem really hard

[Cotra][17:30]

strong upward trend breaks on benchmarks seems like it should be a point toward eliezer verse, even eliezer doesn't want to bet on a specific one?

especially breaks on model size -> perf trends rather than calendar time trends

[Christiano][17:30]

I think that any big break on model size -> perf trends are significant evidence

[Cotra][17:31]

meta-learning working with small models?

e.g. model learning-to-learn video games and then learning a novel one in a couple subjective hours

[Christiano][17:31]

I think algorithmic/architectural changes that improve loss as much as 10x'ing model, for tasks that looking like they at least *should* have lots of economic value

(even if they don't end up having lots of value because of deployment bottlenecks)

is the meta-learning thing an Eliezer prediction?

(before the end-of-days)

[Cotra][17:32]

no but it'd be an anti-bio-anchor positive trend break and eliezer thinks those should happen more than we do

[Christiano][17:32]

fair enough

a lot of these things are about # of times that it happens rather than whether it happens at all

[Cotra][17:32]

yeah

but meta-learning is special as the most plausible long horizon task

[Christiano][17:33]

e.g. maybe in any given important task I expect a single "innovation" that's worth 10x model size? but that it still represents a minority of total time?

hm, AI that can pass a competently administered turing test without being economically valuable?

that's one of the things I think is ruled out before 4 year doubling, though Eliezer probably also doesn't expect it

[Yudkowsky: ]

[Cotra][17:34]

what would this test do to be competently administered? like casual chatbots seem like they have reasonable probability of fooling someone for a few mins now

[Christiano][17:34]

I think giant google-automating-google projects without big external economic impacts

[Cotra][17:34]

would it test knowledge, or just coherence of some kind?

[Christiano][17:35]

it's like a smart-ish human (say +2 stdev at this task) trying to separate out AI from smart-ish human iterating a few times to learn about what works

I mean, the basic ante is that the humans are *trying* to win a turing test, without that I wouldn't even call it a turing test

dunno if any of those are compelling @Eliezer

something that passes a like "are you smart?" test administered by a human for 1h, where they aren't trying to specifically tell if you are AI

just to see if you are as smart as a human

I mean, I guess the biggest giveaway of all would be if there is human-level (on average) AI as judged by us, but there's no foom yet

[Yudkowsky][17:37]

I think we both don't expect that one before the End of Days?

[Christiano][17:37]

or like, no crazy economic impact

I think we both expect that to happen before foom?

but the "on average" is maybe way too rough a thing to define

[Yudkowsky][17:37]

oh, wait, I missed that it wasn't the full Turing Test

[Christiano][17:37]

well, I suggested both

the lamer one is more plausible

[Yudkowsky][17:38]

full Turing Test happeneth not before the End Times, on Eliezer's view, and not before the first 4-year doubling time, on Paul's view, and the first 4-year doubling happeneth not before the End Times, on Eliezer's view, so this one doesn't seem very useful

9.13. GPT-n and small architectural innovations vs. large ones

[Christiano][17:39]

I feel like the biggest subjective thing is that I don't feel like there is a "core of generality" that GPT-3 missing

I just expect it to gracefully glide up to a human-level foaming intelligence

[Yudkowsky][17:39]

the "are you smart?" test seems perhaps passable by GPT-6 or its kin, which I predict to contain at least one major architectural difference over GPT-3 that I could, pre-facto if anyone asked, rate as larger than a different normalization method

but by fooling the humans more than by being smart

[Christiano][17:39]

like I expect GPT-5 would fail if you ask it but take a long time

[Yudkowsky][17:39]

that sure is an underlying difference

[Christiano][17:39]

not sure how to articulate what Eliezer expects to see here though

or like what the difference is

[Cotra][17:39]

something that GPT-5 or 4 shouldn't be able to do, according to eliezer?

where Paul is like "sure it could do that"?

[Christiano][17:40]

I feel like GPT-3 clearly has some kind of "doesn't really get what's going on" energy

and I expect that to go away

well before the end of days

so that it seems like a kind-of-dumb person

[Yudkowsky][17:40]

I expect it to go away before the end of days

but with there having been a big architectural innovation, not Stack More Layers

[Christiano][17:40]

yeah

whereas I expect layer stacking + maybe changing loss (since logprob is too noisy) is sufficient

[Yudkowsky][17:40]

if you name 5 possible architectural innovations I can call them small or large

[Christiano][17:41]

1. replacing transformer attention with DB nearest-neighbor lookup over an even longer context

[Yudkowsky][17:42]

okay 1's a bit borderline

[Christiano][17:42]

2. adding layers that solve optimization problems internally (i.e. the weights and layer N activations define an optimization problem, the layer N+1 solves it) or maybe simulates an ODE

[Yudkowsky][17:42]

if it's 3x longer context, no biggie, if it's 100x longer context, more of a game-changer

2 - big change

[Christiano][17:42]

I'm imagining >100x if you do that

3. universal transformer XL, where you reuse activations from one context in the next context (RNN style) and share weights across layers

[Yudkowsky][17:43]

I do not predict 1 works because it doesn't seem like an architectural change that moves away from what I imagined to be the limits, but it's a big change if it 100xs the window

3 - if it is only that single change and no others, I call it not a large change relative to transformer XL Transformer XL itself however was an example of a large change - it didn't have a large effect but it v what I'd call a large change.

[Christiano][17:45]

4. Internal stochastic actions trained with reinforce

I mean, is mixture of experts or switch another big change?

are we just having big changes non-stop?

[Yudkowsky][17:45]

4 - I don't know if I'm imagining right but it sounds large

[Christiano][17:45]

it sounds from these definitions like the current rate of big changes is > 1/year

[Yudkowsky][17:46]

5 - mixture of experts: as with 1, I'm tempted to call it a small change, but that's because of my mod of it as doing the same thing, not because it isn't in a certain sense a quite large move away from St: More Layers

I mean, it is not very hard to find a big change to try?

finding a big change that works is much harder

[Christiano][17:46]

several of these are improvements

[Yudkowsky][17:47]

one gets a minor improvement from a big change rather more often than a big improvement from a l change

that's why dinosaurs didn't foom

[Christiano][17:47]

like transformer -> MoE -> switch transformer is about as big an improvement as LSTM vs transforme

so if we all agree that big changes are happening multiple times per year, then I guess that's not the difference in prediction

is it about the size of gains from individual changes or something?

or maybe: if you take the scaling laws for transformers, are the models with impact X "on trend," with changes just keeping up or maybe buying you 1-2 oom of compute, or are they radically better / scal much better?

that actually feels most fundamental

[Yudkowsky][17:49]

I had not heard that transformer -> switch transformer was as large an improvement as lstm -> transformers after a year or two, though maybe you're referring to a claimed 3x improvement and comparing that to the claim that if you optimize LSTMs as hard as transformers they come within 3x have not examined these claims in detail, they sound a bit against my prior, and I am a bit skeptical both of them)

so remember that from my perspective, I am fighting an adverse selection process and the Law of Earlier Success

[Christiano][17:50]

I think it's actually somewhat smaller

[Yudkowsky][17:51]

if you treat GPT-3 as a fixed thingy and imagine scaling it in the most straightforward possible way, tl I have a model of what's going on in there and I don't think that most direct possible way of scaling g you past GPT-3 lacking a deep core

somebody can come up and go, "well, what about this change that nobody tried yet?" and I can be li "ehhh, that particular change does not get at what I suspect the issues are"

[Christiano][17:52]

I feel like the framing is: paul says that something is possible with "stack more layers" and eliezer isr We both agree that you can't literally stack more layers and have to sometimes make tweaks, and al that you will scale faster if you make big changes. But it seems like for Paul that means (i) changes to stay on the old trend line, (ii) changes that trade off against modest amounts of compute

so maybe we can talk about that?

[Yudkowsky][17:52]

when it comes to predicting what happens in 2 years, I'm not just up against people trying a broad range of changes that I can't foresee in detail, I'm also up against a Goodhart's Curse on the answer being a weird trick that worked better than I would've expected in advance

[Christiano][17:52]

but then it seems like we may just not know, e.g. if we were talking LSTM vs transformer, no one is going to run experiments with the well-tuned LSTM because it's still just worse than a transformer (though they've run enough experiments to know how important tuning is, and the brittleness is much of why one likes it)

[Yudkowsky][17:53]

I would not have predicted Transformers to be a huge deal if somebody described them to me in advance of having ever tried it out. I think that's because predicting the future is hard not because I'm especially stupid.

[Christiano][17:53]

I don't feel like anyone could predict that being a big deal

but I do think you could predict "there will be some changes that improve stability / make models slightly better"

(I mean, I don't feel like any of the actual humans on earth could have, some hypothetical person co

[Yudkowsky][17:57]

whereas what I'm trying to predict is more like "GPT-5 in order to start-to-awaken needs a change via which it, in some sense, can do a different thing, that is more different than the jump from GPT-1 to GPT-3; and examples of things with new components in them abound in Deepmind, like Alpha Zero having not the same architecture as the original AlphaGo; but at the same time I'm also trying to account for being up against this very adversarial setup where a weird trick that works much better than I expected may be the thing that makes GPT-5 able to do a different thing"

this may seem Paul-unfairish because any random innovations that come along, including big changes that cause small improvements, would tend to be swept up into GPT-5 even if they made no more difference than the whole thing with MoE

so it's hard to bet on

but I also don't feel like it - totally lacks Eliezer-vs-Paul-ness if you let yourself sort of relax about that and just looked at it?

also I'm kind of running out of energy, sorry

[Christiano][18:03]

I think we should be able to get something here eventually

seems good to break though

that was a lot of arguing for one day

You are probably underestimating how good self-love can be

I am very grateful to the following people, in general, and for their helpful feedback on this post: Nick Cammarata, Kaj Sotala, Miranda Dixon-Luinenburg, Sam Clarke, Mrinank Sharma, Matej Vrzala, Vlad Firoiu, Ollie Bray, Alan Taylor, Max Heitmann, Rose Hadshar, and Michelle Hutchinson.

I was on a plane to Malta when I realised I had lost something precious. I was struggling to meditate. I knew there was some disposition that made meditation easier for me in the past, something to do with internal harmony and compassion and affection. Alas, these handles failed to impact me. On a whim, I decided to read and meditate on some of my notes. 3h later, I had recovered the precious thing. It was one of the most special experiences of my life. I felt massive relief, but I was also a little scared--I knew that this state would likely pass. I made a promise to myself to not forget what I felt like, then, and to live from that place more. This post is, in part, an attempt to honour that promise. I spent most of my holiday in Malta reading about and meditating on the precious thing, and I now feel like I'm in a place where I can share something useful.

This post is about self-love. Until recently, I didn't know that self-love was something I could aim for; that it was something *worth* aiming for. My guess is that I thought of self-love as something vaguely Good, a bit boring, a bit of a chore, a bit projection-loaded (I'm lovable; I love me so you can love me too), and lumped together with self-care (e.g. taking a bath). Then I found [Nick Cammarata](#) on Twitter and was blown away by the experiences he was describing. Nick tweeted about self-love from Sep 2020 to May 2021, and then moved on to other things. His is the main body of work related to self-love that I'm aware of, and I don't want it to be lost to time. My main intention with this post is to summarise Nick's work and build on it with my experiences; I want to get the word out on self-love, so that you can figure out whether it's something *you* want to aim for. But I'm also going to talk a little about how to cultivate it and the potential risks to doing that. One caveat to get out of the way is that I'm a beginner--I've been doing this stuff for under a year, for way less than 1h/day. Another is that I expect that my positive experiences with self-love are strongly linked to me being moderately depressed before I started.

What is self-love?

Self-love is related to a lot of things and I'm not sure which are central. But I can point to some experiences that I have when I'm in high self-love states. While my baseline for well-being and self-love is significantly higher than it used to be, and I can mostly access self-love states when I want to, most of the time I am not in very high self-love states, because my attention is elsewhere. Some of the following experiences point to the core of what self-love feels like, some are actions or tendencies that self-love spins up out of, and some are consequences of self-love. It is hard to untangle these categories so I don't try to.

- Take a second to imagine the love you might feel towards a newborn child or a cute animal. They probably haven't done anything to 'earn' your love; they

might even be acting unskillfully (admittedly, I don't know what a skilful baby looks like). But you might love them anyway. Self-love feels quite like that for me: unconditional, newborn love.

- I feel bad *about myself* a lot less: I'll notice a character defect or a way I acted unskillfully, and won't feel bad *about myself*-similarly to how I would feel about a close friend messing up. It doesn't follow from this that I don't feel bad (I think traditionally 'negative' emotions can be functional), don't want to change, or act differently in the future. More on this later. A consequence of this is that it is easier to see my imperfections, and to see the world, as opposed to flinching away from them. *When states of the world directly impact your perceived self-worth, it can be really scary to see the world as it is.* Some examples: a [kid](#) who wants to be a writer and cannot admit that she made a spelling mistake; my aversion to studying AI safety because doing that puts me in contact with the fact that I don't know that much and hence that I'm worthless.
- Affection: I'll drop and smash a plate, and where previously there might have been some frustration or self-judgement, the mental motion might be "Oh, silly Charlie, I still love you". Or when I toned down a claim in this post just now I was like "Oh thank you for protecting me". Importantly, I'm not saying empty words--I'm translating my feelings into words. The examples of affection above closely resemble how I'd feel towards a small child, but the affection can also feel more friend or partner-like. For example, I got drunk for the first time in a while last weekend, and found drunk Charlie really adorable.
- Compassionate awareness: I'll define "compassion" as seeing suffering and being moved by it, where "being moved" might connote warmth and caring and non-judgement and desire to help. I'm often including my experience (emotions, thoughts, sensations) in my moment-to-moment awareness, greeting and feeling what's happening to me. Sometimes I'll notice that I'm conflicted or struggling or suffering, and will dive deeper. I find it useful to view myself as having many [parts](#), who have different feelings and goals and functions. So I'll often be talking to my parts and figuring out what they want and why, how they're feeling, what they think of each other--and holding compassionate space for all of that to happen in. I find the parts model pretty useful for compassion and affection, in part because it's easier for me to feel/send love when there's some distance between the lover and the loved.
- Nick Cammarata says that a heuristic for self-love is that you feel like you're walking around with someone you have a crush on ([here](#) is a thread from him about this, and [here](#) is one where he discusses the controversy that thread caused). It feels like that for me: romantic. But read "romantic" more as beautiful and exciting than including desire or projection, if those are part of your dating experience. There's curiosity--wanting to know more about my experience--and awe and affection, and joy, at being able to share these moments with myself.
- Relatedly, I feel like I'm "with myself", as opposed to "by myself". This is how Nick describes feeling too: "My body feels different. Being in my body used to feel a bit like being in a neutrally-charged hollow shell interacting with the world, now it feels a bit like a stable and warm castle with a cozy quality. I feel like I am "with myself" inside of it. Others are outside, and I can open the castle and feel close to them, but staying inside with myself is the default."
- Loving action: For example, attending to my experiences; paying attention to what I want and acting to make that happen; prioritising resolving internal conflict; not ignoring or shutting parts of me down; noticing a flicker of not-ok-ness while watching TV and pausing the TV to figure out what's wrong and whether it's ok to continue watching.

- I feel substantially safer, like I have a blanket wrapped around me. I don't fully understand this, but I think it's because I'm clinging less to external conditions being satisfied in order to feel worthwhile. I feel like I'm more capable of taking whatever the world throws at me. I still care about the external things, like whether a partner loves me, but I don't [cling](#) to them in the same way.
- Worthiness/self-esteem: I used to have a strongly bad filter on my self-perception. Now I can more easily remember (to some extent) my inherent goodness and preciousness and beauty. I can also more clearly see all of the amazing things about me.
- Spending time with myself used to be unbearable--I would sink into awareness-collapsing distractions. In these states, spending time with myself is really fun, often more fun than spending time with friends. Time alone is nourishing and special.
- Happiness: In my experience, it is a lot easier to do anything when I have surplus happiness, and it is extremely difficult to do anything when depressed. Self-love makes me very happy so I'm able to do the things that matter to me. [This](#) is a tweet thread where Nick writes that raising happiness baselines is possible and incredibly important.
- Energy: Part of this is fighting myself less, which includes less suffering-based motivation and internal conflict. Freeing up those resources has been astonishingly powerful for me. Part is not needing to invest emotional resources into trying and needing to feel loved, because I have a wellspring within.
- More love for others and the world.

But won't I turn into jello?

It's easy to imagine that, if you feel unconditionally worthwhile, if you have access to a deep source of self-compassion and affection and joy, then you will care less about changing or pursuing your goals. This was my worry, so I want to tackle it head-on.

I think turning into jello is a very understandable worry. A lot of people go their whole lives making their self-worth conditional in order to act better: they take damage--dislike or judge themselves--whenever they act imperfectly or realise they are imperfect or don't achieve the things they want to. In a world as unfair and uncontrollable as this one, I think taking so much damage is often not that functional. Moreover, I claim that *you can care deeply while feeling worthwhile and suffused with compassion and affection and joy*. All that said, messing with the strategy that helps you act better is a big deal (see Risks).

I don't have any good arguments about how often we'd expect people to turn to jello, besides looking at the people who have walked the path. However, I'm confident that more self-love does not *necessitate* less caring, because I and many others have experienced that more self-love leads to *more* caring. Nick Cammarata says that he has never seen people turn into jello, and that, "In fact, it usually pushes [people] far in the other direction". This accords with the behaviourism literature (at least as summarized in "[Don't Shoot the Dog](#)"), which claims that both animals and humans are best trained by only giving them rewards and no punishments. This probably generalizes to internal rewards and punishments, which are largely learned and internalized based on how people have treated us in the past.

I'm reminded of Nate Soares' writings on [Replacing Guilt](#). He writes that it's you that cares about your goals, that wants to become stronger or save the world. Those things that you actually care about won't go away with more self-love; what changes is your strategy for pursuing them. **You no longer pursue things in order to feel worthwhile, but simply because you want to.** Indeed, it is not self-loving to shut down those parts of you that care about things. An essential component of self-love, as I see it, is being there with and feeling fully whatever is happening for me, especially when I want things to be different.

Risks

- Your goals and strategies might change, even if your values remain the same. For example, I was aiming to pursue a PhD in machine learning, partly because I thought it would make me worthwhile. When I felt worthwhile I stopped that; I was able to think more freely about which strategy looked best according to my values.
- Changing your brain might have negative effects in the short term, even if things are good in the end. For example, as I walked down the self-love path I felt my external obligations start to drop away. While things are clearly better now, I'm still figuring out how to be internally motivated and also get shit done, and for a while I got less shit done than when I was able to coerce myself.
- It's easy to misunderstand what you're aiming for.
- It's also easy to miss your target. For example, for a while, I used to throw what I thought was compassion at negative emotions and they would go away. And on the surface that seems kinda reasonable. But, for me, industry-grade compassion requires seeing the emotions fully--holding them and understanding them and letting them be felt as strongly as they want to be felt.
- The techniques you use to develop self-love might have side effects of their own. For example, if you're doing a lot of meditation I would be surprised if you didn't have some negative experiences at some point. (That said, lovingkindness meditation seems like one of the safer types, and I'm broadly pro-meditation.)

How to self-love

I'm really confused about this, sorry. The path is muddy, at least to me. That's why I focused on describing self-love. I realise that this might be frustrating, especially if I managed to get you excited about self-love. That said, I decided to write *something* here rather than nothing. Please take this section with a bunch of salt.

Nick thinks that the two most promising avenues are solo MDMA trips and metta (lovingkindness) meditation.

MDMA: I am not recommending that people take MDMA, because that would be illegal, and because I have no idea what your situation is. If you intend to take MDMA, please do some research on safety (e.g. read at least [this](#) and [this](#)) to get a sense of the costs, and because you can substantially reduce risks and side effects if you do decide to take it. Here is my impression of the benefits: MDMA makes you feel a lot of love--very likely a *lot* more than you've ever experienced, possibly orders of magnitude more--including self-love. I've seen and heard of many people experiencing

extremely large and lasting improvements to self-love when they take MDMA alone, close their eyes, and focus on investigating their experiences--including how they relate to themselves. This accords with preliminary [research](#) on the efficacy of MDMA-assisted therapy for PTSD. My guess for why this happens is that MDMA is extremely good at memory reconsolidation *a la* [Unlocking the Emotional Brain](#), presumably because it makes painful experiences/memories safe to look at and the love makes them easy to rewrite. Another benefit is that it gives you some information about what it's like to have self-love--for example, you might for the first time experience complete self-acceptance while also caring deeply about doing things and changing, and that's cool if that was a crux for you wanting more self-love. It also substantially clarifies what to aim for when sober, which is important:

Nick: "I can't overstate how impossible it would have been for me to get to a state of self-love without MDMA, even after hundreds of hours of metta (which I did before the MDMA). Not sure it'll be the case for everyone, but I suspect it makes things way easier".

Lovingkindness meditation (metta): Metta is a slower way to increase your capacity for love, albeit substantially. You could think of metta as doing reps to strengthen the love muscle (but more beautiful than that). Alongside love, metta also builds awareness, [indistractability, sensory clarity, and equanimity](#)--all of which are pretty useful for self-love. Below, I discuss some introspection techniques that might be useful. One reason to expect metta to be better than those techniques is that, when you're good at it, metta has feedback loops that can get you into very high self-love states. **Resources:** The canonical metta book is [this](#) one, but I think it's mostly good for giving you models and not for practice. Kaj Sotala says that lots of people find [TWIM](#) really effective. [Here's](#) a guided meditation and [here's](#) one with a different style. You can do [concentration meditation](#) with love as the object of concentration instead of the breath, and can get coaching for that [here](#).

Other things that might help

I wrote this section for someone like me a year ago--someone who strongly wants self-love and is desperate to read anything they can about how to get it. Consequently, this section is long and lower-quality; feel free to skip it!

Deepen your understanding of self-love: If you have an hour or two, [searching](#) for '@nickcammarata self-love' might be the best use of your self-love time. You could also try to spend lots of time with/read/take workshops with/take retreats with people who are really good at self-love. I don't know of specific people but you might find some within Nick's Twitter circles and (lovingkindness/metta) meditation communities (Tara Brach, Sharon Salzburg).

Be with yourself: Being with yourself (your experiences) is the training ground for self-love. It is hard to become your best friend if you do not know yourself. Being with yourself requires some baseline self-love, though--it might not be good or advisable at first. One idea is to walk around without external input when possible. You can also be with yourself whenever you notice suffering, or even moment-to-moment (e.g. while working)--though this requires some skill to be able to do with little cost (see [this](#) course). I refresh my awareness about my experiences very frequently, and sometimes the awareness is roughly continuous.

Figure out what you believe: The ability to self-love seems strongly mediated by ones (implicit) beliefs about whether it's safe and good to do so. So I would focus on figuring out what you believe. Indeed, many of the introspection techniques I list below work to facilitate this process, and could be done with this process in mind. You can ask yourself, with gentle curiosity, why you don't want self-love or why you think it's good to make your self-worth conditional. This is important because there are probably reasons, and your current set-up might be doing something very useful (such as [guilt-based motivation](#)). And until you understand those functions it will be hard and maybe bad to shift things up.

Introspection/therapy techniques: Explaining each technique well is beyond the scope of this post, but I have linked to a short blog post and a more comprehensive resource where possible.

- [Coherence therapy \(book\)](#) and [memory reconsolidation for self-affection](#). I think these posts are worth reading even if you don't intend to practice the techniques, because they have useful models of how therapy progress works.
- [Focusing \(book\)](#) is a bread-and-butter self-inquiry technique. Getting good at focusing will probably aid many self-love strategies. You can also use focusing to enquire directly about self-love. [Jack](#) from CFAR facilitates high-quality and very reasonably-priced focusing sessions.
- [Internal family systems \(book\)](#) is a type of therapy that works with your parts. You could try to find an IFS therapist but I expect the average therapist to be bad.
- I haven't tried this but my therapist (who I trust) recommends [compassion-focused therapy](#). "The primary focus of CFT is identifying sources of resistance to (self-)compassion and then building and strengthening the compassionate self (the Healthy Adult mode in Schema Therapy; the Wise Mind in Dialectical Behavior Therapy)."
- I just heard about [Core Transformation \(book\)](#) and it seems really cool but I don't know how cool. Maybe read the [blog post](#)?
- Kaj Sotala says he got significant value from guided Ideal Parent Figure practice ([guided meditation](#), [course](#), [book](#)--see chapter 8). The idea is that a lot of our emotional conditioning around self-worth comes from childhood, where we learn what kinds of behaviours get us love and acceptance from our caregivers. IPPF exploits the fact that the emotional brain doesn't fully distinguish between the real and the imagined, so you can reprogram your mind by imagining yourself as a young child with ideal parents who always express unconditional love and delight towards you.
- The [exercises](#) from [Self-Compassion \(book\)](#) seem pretty good. I read the book a while ago but can't remember how good it was.

Miscellaneous:

- Other meditation: either [concentration](#) (this is mostly what I did) or [noting](#) (what I weakly recommend now). These techniques build skills ([concentration](#), [sensory clarity](#), [equanimity](#), and awareness) that I expect to indirectly affect all of your other self-love endeavors (including MDMA). You can get coaching for concentration (and metta) meditation [here](#).
- [Expanded awareness](#): A big part of self-love is holding your experience in awareness. This course will help you do that.
- Therapy.

Other books:

- [Radical Acceptance](#) (I remember really enjoying this). It worked strongly on the belief level for me. Some people might be a bit allergic to the Buddhist stuff.
- [Replacing Guilt](#) (is truly awesome). Also works really strongly on the belief level. Written for people like you.
- [Mindful Compassion](#) (my therapist recommends this and I trust her but I haven't read it).

Transcript: "You Should Read HPMOR"

The following is the script of a talk I gave for some current computer science students at my alma mater, [Grinnell College](#). This talk answers "What do I wish I had known while at Grinnell?".

Hi, I'm Alex Turner. I'm honored to be here under Sam's invitation. I'm in the class of 2016. I miss Grinnell, but I miss my friends more—enjoy the time you have left together.

I'm going to give you the advice I would have given Alex₂₀₁₂. For some of you, this advice won't resonate, and I think that's OK. People are complicated, and I don't even know most of you. I don't pretend to have a magic tip that will benefit everyone here. But if I can make a big difference for one or two of you, I'll be happy.

I'm going to state my advice now. It's going to sound silly.

You should read a Harry Potter fanfiction called [Harry Potter and the Methods of Rationality](#) (HPMOR).

I'm serious. The intended benefits can be gained in other ways, but HPMOR is the best way I know of. Let me explain.

When I was younger, I was operating under some kind of haze, a veil, distancing me from what I really would care about.

I responded to social customs and pressure, instead of figuring out what is good and right by my own lights, how to make that happen, and then executing. Usually it's fine to just follow social expectations. But there are key moments in life where it's important to reason on your own.

At Grinnell, I exemplified a lot of values I now look down on. I was extremely motivated to do foolish or irrelevant things. I fought bravely for worthless side pursuits. I don't even like driving, but I thought I wanted a fancy car. I was trapped in my own delusions because I wasn't thinking properly.

Why did this happen, and what do I think has changed?

On Caring

First, I was disconnected from what I would have really cared about upon honest, unflinching reflection. I thought I wanted Impressive Material Things. I thought I wanted a Respectable Life. I didn't care about the bible, but I brought it with me to my dorm anyways so that I'd be more "wholesome" according to my cultural background. Chasing something someone convinced me to believe I wanted, but which I don't care about.

I became motivated to *unironically reflect on what is good, how I want the universe to look by the time I'm done with it—to reason about what matters without asking for permission*. Not so that you can show how caring you are on social media. But

because some things are *fucking important*. Peace, learning, freedom, health, justice. Human flourishing. Happiness.

When I inhabit my old ways of thinking about altruism, they evoke guilt and concern: “The world will burn. I have to do my part.” If, however, I’ve discharged my duties by donating and recycling and such, then I no longer feel guilty. But the cruel fact is that *no matter what I do, millions of people will die of starvation this year*. Due to a coincidence of space and time, none of these people happen to be my brother or sister, my mother or father. None are starving two feet away from me. But who cares if someone starves two feet away, or 42 million feet away—they’re still starving!

What I’m saying here is subtler than “care a lot.” I’m gesturing at a particular *kind* of caring. The kind from the [assigned essay](#). Since you all read it, I probably don’t need to explain further, but I will anyways. Some extreme altruists give almost everything they have to charity. It’s natural to assume they have stronger “caring” feelings than you do, but that may not be true.

The truth is that I am biologically incapable of caring as much as *9 million x (how much I would care if my brother starved)*. My internal “caring system” doesn’t go up that many decibels, it just silently throws an emotion overflow error. Does that mean I can’t, or don’t want to, dedicate my life to altruism? No. It means I ignore my uncalibrated emotions, that I do some math and science to estimate how I can make the biggest difference, and then do that.

What does this have to do with Harry Potter? HPMOR made me realize I *should* care in this way. HPMOR let me experience *the point of view of someone intelligently optimizing the world to be a better, more moral place*. HPMOR let me look through the eyes of someone who deeply cares about the world and who tries to do the most good that they can. The *experience* counts.

You’ll notice that CS-151 doesn’t start off with a [category theoretic-motivation of functional programming](#) in Scheme, with armchair theorizing about loop invariants, parametric polymorphism, and time complexity. There are labs. You experience it yourself. That’s how the beauty of computer science *sticks to you*.

HPMOR is the closest thing I know to a lived experience of *gut-level caring about hammering the world into better shape*.

On Foolishness

Second, in 2016, I was enthusiastic, optimistic, and hard-working. I was willing to swim against social convention. I was also foolish.

By “foolish”, I don’t quite mean “I did pointless things.” I mean: “My cognitive algorithm was not very good, and so I did pointless things.” By analogy, suppose it’s 2026, and you’re doing research with the assistance of a machine learning model. Given a hypothesis, the model goes off and finds evidence. But suppose that the model *anchors* on the first evidence it finds: Some study supports the idea, and then the model selectively looks for *more evidence for its existing beliefs!* Wouldn’t this just be *so annoying and stupid?*

In some parts of your life, you are like this. Yes, *you*. Our brains regularly make embarrassing, biased mistakes. For example, I stayed in a relationship for a year too

long because I was not honest with myself about how I felt.

In 2014, I scrolled past a News Feed article in which Elon Musk worried about extinction from AI. I rolled my eyes—“Elon, AI is great, you have no idea what you’re talking about.” And so I kept scrolling. (If someone made a biopic about me, this is where the canned laugh track would play.)

The mistake was that I had a strong, knee-jerk opinion about something I’d never even thought about. In 2018, I reconsidered the topic. I ignored the news articles and sought out the best arguments from each side of the debate. I concluded that my first impression was *totally, confidently wrong*. What an easy way to waste four years. I’m now finishing my dissertation on reducing extinction risk from AI, [publishing papers](#) in top AI conferences.

My cognitive algorithm was not that great, and so I made many costly mistakes. Now I make fewer.

What, pray tell, does this have to do with Harry Potter? HPMOR channels someone who tries to improve their thinking with the power and insight granted by behavioral economics and cognitive psychology, all in pursuit of worthy goals. The book gave me a sense that *more is possible*, in a way that seems hard to pick up from a textbook. (I took cog-psych classes at Grinnell. They were evidently insufficient for this purpose: I didn’t even realize that I should try to do better!)

HPMOR demonstrates altruistic fierceness: How can I make the future *as bright as possible*? How can I make the *best* out of my current situation? What kinds of thinking help me arrive at the truth *as quickly as possible*? What do I think I know, and why do I think I know it? What would reality look like if my most cherished beliefs were *wrong*?

In the real world, we may stand at the relative beginning of a bright and long human history. But to ensure that humanity *has* a future, to make things go *right*—that may require finding the truth as quickly as possible. That may require clever schemes for doing the *most* good we can, whatever we can. That may require altruistic fierceness. (See: the [Effective Altruism](#) movement.)

Taken together, *caring deeply about maximizing human fulfillment* and *improving my cognitive algorithms* changed my life. I don’t know if this particular book will have this particular effect on you. For example, you might not be primarily altruistically motivated on reflection. That’s fine. I think you may still selfishly benefit from this viewpoint and skillset.

HPMOR isn’t the only way to win these benefits. But I think it’s quite good for some people, which should make it worth your time to try 5–10 chapters. I hope you benefit as much as I did.

You can find the book at www.hpmor.com (I recommend the [PDF version](#)). You can find the unofficial, very good podcast reading [on Spotify](#). You can find me at turneale@oregonstate.edu.

Where did the 5 micron number come from? Nowhere good. [Wired.com]

This is a linkpost for <https://www.wired.com/story/the-teeny-tiny-scientific-screwup-that-helped-covid-kill/>

This article describes a scientist's attempt to figure out where the 5 micron number, and general belief that most respiratory diseases weren't airborne, came from. She eventually traces it back to a particular number developed for a very different purpose.

I have not fact checked it extensively, but last winter I did try to look into the general state of knowledge on airborne transmission vs fomites and found it weirdly empty, in ways that are consistent with this article.

Money Stuff

Cross-posted from [Putanumonit](#).

Vodka and War

When I was 16 I made the dumbest financial investment of my life. Our family came to live in the US for a few months, and I made \$500 giving tennis lessons to the local kids. My individually directed consumption at the time consisted mostly of video games, food, and booze. And so I faced a decision: should I spend this windfall on 4 years of subscription to EverQuest? 166 falafel sandwiches? 104 bottles of Keglevich flavored vodka at the old central bus station in Tel Aviv upon my return?

Despite these immensely appealing options, somehow my dad convinced me that the adult™ and responsible™ thing to do was to invest the money in a yield-bearing instrument. And since at 16 I liked thinking of myself as a responsible adult™ almost as much as I liked cheap vodka and MMOs, I found myself walking out of the local bank branch holding a purchase certificate for a 10-year US treasury bond.

That's right. Upon coming into possession of more money than I've ever held, I lent it out to George W. Bush at ~0% real rate of return. This was right around the invasion of Iraq, which ultimately cost \$1.9 trillion over 8 years. And so instead of using the \$500 for my own betterment (or, at least, enjoyment), I ended up funding exactly 66 *milliseconds* of the Iraq War. By the time I redeemed the bond a decade later I was close to graduating business school with a job in finance, and the \$1/month it accumulated for the duration didn't move the needle much for my personal financial situation.

What should I have spent it on? Even if we ignore all consideration of fun and personal taste and focus on the money only, the wisest purely financial investment I could have made with that money was, in fact, vodka. I should have come back to Israel and sponsored a rager for all my high school friends and acquaintances. A party would have transformed my reputation from "weirdo who's good at math" to "cool and generous weirdo who's good at math". Given that my high school classmates went on to become influential professionals, businesspeople, and founders of successful companies, earning their respect with free drinks would surely be worth orders of magnitude more a decade later than \$650.

The Value of Friendship

It's strange, or even profane, to think of friendships in terms of their financial value. If we did, we would likely find ourselves struggling for both money *and* friends. But it's also true that one's friend group has a huge impact on one's finances. In a lot of communities, this impact is hugely negative.

On one side are groups united around communist-ish ideology (for whom enforcing equality of broke-assedness is at least thematically appropriate). People in those groups valorize poverty and demonize riches, so any member getting ahead financially is scorned and resented. The richer friends in those groups are often aggressively mooched off of ("from everyone according to their ability!") and then aggressively scapegoated to resolve the cognitive dissonance ("they didn't deserve their job anyway, and their parents are rich besides!"). As long as people aren't starving, making money is often not worth to them the social cost they would incur.

There are also groups entirely capitulated to capitalism, egging each other on in contests of conspicuous consumption. I talked to an Uber driver who confided that he and his friends spend slightly more than all of their disposable income on fashion sneakers. He said that anyone discovered to be prudently managing their money instead of keeping up with the latest sneaker offerings is made fun of. He wasn't particularly happy to be driving 10 hours a day just to fill his small apartment with hideous [Yeezy Foam Runners](#), but didn't really see a way out.

My nerd friends, however, can't tell a Yeezy from a yuzu but instead encouraged each other to buy Bitcoin [as early as 2011](#) which made a large (and growing) number of them rich. The community in general is quite remarkable for having people several orders of magnitude apart in income coexisting quite happily wearing the same t-shirts with programming jokes to the same nerdy meetups and being encouraging both of those who make millions and of those who drop out to focus on non-monetary pursuits. There's also a budding culture of patronage, as exemplified in [community-sourced prizes](#) for important work, a norm of posting bounties for tasks and hiring friends, and the respect granted to Effective Altruism donors.

Bottom line: your friend group affects your finances not just in direct ways like job offers but in shaping how people think about money, how they earn and spend it, and how group status interacts with wealth and income. It's a lot easier to be financially comfortable in a community that's relatively sane about money. Unfortunately, in the realm of personal finance, sanity is very much the rare exception.

Dollar Derangement

People can be deranged about money as a result of their friend group, but people also manage surprising levels of financial insanity all on their own.

I've seen successful entrepreneurs and professionals so averse to dealing with personal finance they'd procrastinate for months on basic tasks like setting up a payment for a loan, watching a stack of unopened envelopes from the bank pile up on their desk. I've seen people get literal anxiety attacks around the tax deadline, the one time a year the threat of jail forced them to peer into the money abyss.

I've seen a professional with a \$500,000 salary who didn't know if she spent \$200,000 or \$800,000 the previous year. I've seen another careful budgeter who had multiple years of runway saved up stay at a job he hated because he couldn't bear the thought of a month with no income.

I've seen people with seven-figure net worth stake it all on a speculative ICO or angel investment. I've seen people with \$100 to invest spend months researching how to earn an extra 0.2% in some crypto yield-farming scheme, "earning" a few cents per hour of labor.

I've seen people on the verge of homelessness be too proud to ask their family or friends for a small loan to tide them through a crisis. I've seen others who were doing quite well who still demanded that their parents pay for their expenses.

I've seen a classmate's family [burn \\$8,000 a year](#) just to avoid talking about lending each other money. I've seen a friend talk about [the complex emotions he dealt with](#) upon receiving a large gift from his parents, and I've seen [thousands of people](#) telling him he's a piece of shit for daring to mention it and in the same breath asking him for charity.

I've seen [turkeys in the jungle](#).

I think I've maintained roughly the same phlegmatic attitude about money as my circumstances changed from being financially dependent, to independent and broke, to employed and in debt, to comfortable. To me, money is instrumental — [a resource to put a](#)

[number on, trade off against time and effort and risk, and manage with basic math](#). But to other people it seems to be a measure of their character, a core part of their identity, and a leading cause of their derangement.

American culture demands total hypocrisy about money at all times. If you're broke you must pretend to have money so people don't judge you, but if you have a lot of money you must pretend to have less for the same reason — unless you're very rich in which case you pretend to be much richer than you are yet again. If you don't care about money and slack off you're a bum, if you are motivated by money you must pretend (even to your employer) that you're not, but you must also act as if everyone else is motivated *only* by money or you're naïve. Everyone agrees that money should be divorced from morals except everyone believes either that rich people are immoral because they're greedy materialists, that poor people are immoral because they're lazy moochers, or [both at the same time](#).

This isn't just a Rationalist complaint about some part of social reality being less than perfectly legible. The problem is that money, which is measurable and legible, gets tangled up in the parts of social reality that are anything but, like status and tribal affiliation. The irreconcilable contradictions that result from this derange people's brains, their bank accounts, and national politics.

Redistributions

Here's [anthropologist Xiang Biao](#) talking about what happens when money and social status become too entangled:

From an anthropological viewpoint, [China] is something of a special case. Does competition exist in other societies, especially primitive societies? Yes, but there are a couple of things worth noting. The first is that people's lives are often made up of two parts: the sphere of prestige and the sphere of subsistence. Subsistence refers to hunting and farming, in which people usually cooperate rather than compete so that everyone can be fed. However, competition still exists in this kind of society — usually between leaders and heads of tribes or families. They're typically male, and they have competitive relationships with leaders in other villages. What are they competing for? Prestige. So, competition exists when it comes to prestige. Perhaps the best-known example is potlatch, in which these tribal heads compete for prestige by distributing their accumulated wealth with others, or by destroying their own wealth in public. This is interesting, because this vying for prestige is directly related to redistribution. Leaders get prestige by sharing their wealth, so competing for prestige is done through material redistribution which then contributes to equality.

In terms of subsistence, however, there is no competition. This is where things can be complicated. For example, everyone has different abilities at hunting. Let's say you are skilled and killed a deer. Everyone will recognize you in terms of prestige, and they will praise your courage and hunting skills. But the meat must be evenly distributed. In China's case, this kind of differentiation no longer exists. Competition is total.

In wealthy societies like the US, **people are aggressively trying to redistribute prestige to themselves while talking only about redistributing sustenance to others**. This is the root cause of a lot of political dysfunction and mutual hatred.

The number of people in rich nations struggling for basic material sustenance like food and shelter is quite small, but not zero. There are just enough of them to serve as the motte for the prestige hunters. In any case, the homeless and those working three shifts aren't the ones writing Twitter threads and magazine articles denouncing rich people, while the activists and journalists who do are rarely destitute or welfare-dependent.

Almost everyone is, however, struggling for status and respect. Sustenance itself requires a perceived sacrifice of dignity, whether it means you have to submit to a boss or wrangle with an impersonal bureaucracy or plead with your family. When people do acquire disposable income, they tend to spend it on status competitions like limited spots in selective clubs and schools for their kids. People who think they've opted out of the status rat race to settle in material comfort [still find prestige games biting them in the ass](#).

It's less acceptable to say "let's take this person's status and redistribute it among ourselves" so people conflate prestige with money and pretend that rich people are hoarding *both* in some immoral way. As my friend says, this makes billionaires a universally accepted target of hate and [the most oppressed class in America](#).

Almost no one (outside of Effective Altruism, perhaps) is offering to give billionaires more prestige *in exchange* for their money. Warren Buffett is widely admired for making billions of dollars, not for [giving 99% of them away](#). Mark Zuckerberg got more goodwill for surfing with a flag than for donating a 12-figure sum to public schools.



So: the chattering classes are looking to appropriate the elite's prestige by pretending to care about redistributing money to the poor. Rich people compete with each other for prestige mostly by acquiring more money, and have no reason to give any of it away if to those who would take their prestige along with it. The actual poor lack the platform and the leisure for these status fights, and their voice is mostly drowned out by the rest.

This situation contributes to economic inequality by sidelining discussions of effective redistribution, makes everyone polarized and hate each other, and feeds back into Americans' individual insanity around money.



Jakeup
@yashkaf



Food \$200

Data \$150

Rent \$800

Negative-sum status competitions \$3,600

Utility \$150

someone who is good at the economy please help me
budget this. my family is dying

7:10 PM · Jul 25, 2021



How do we become confident in the safety of a machine learning system?

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

Thanks to Rohin Shah, Ajeya Cotra, Richard Ngo, Paul Christiano, Jon Uesato, Kate Woolverton, Beth Barnes, and William Saunders for helpful comments and feedback.

Evaluating proposals for building safe advanced AI—and actually building any degree of confidence in their safety or lack thereof—is extremely difficult. Previously, in “[An overview of 11 proposals for building safe advanced AI](#),” I tried evaluating such proposals on the axes of outer alignment, inner alignment, training competitiveness, and performance competitiveness. While I think that those criteria were good for posing open questions, they didn’t lend themselves well to actually helping us understand what assumptions needed to hold for any particular proposal to work. Furthermore, if you’ve read that paper/post, you’ll notice that those evaluation criteria don’t even work for some of the proposals on that list, most notably [Microscope AI](#) and [STEM AI](#), which aren’t trying to be outer aligned and don’t really have a coherent notion of inner alignment either.

Thus, I think we need a better alternative for evaluating such proposals—and actually helping us figure out what needs to be true for us to be confident in them—and I want to try to offer it in the form of *training stories*. My hope is that training stories will provide:

- a general framework through which we can evaluate any proposal for building safe advanced AI,
- a concise description of exactly what needs to be true for any particular proposal to succeed—and thus what we need to know to be confident in it—and
- a well-defined picture of the full space of possible proposals, helping us think more broadly regarding new approaches to AI safety, unconstrained by an evaluation framework that implicitly rules out certain approaches.

What's a training story?

When you train a neural network, you don’t have direct control over what algorithm that network ends up implementing. You do get to incentivize it to have some particular behavior over the training data, so you might say “whatever algorithm it’s implementing, it has to be one that’s good at predicting webtext”—but that doesn’t tell you *how* your model is going to go about accomplishing that task. But exactly how your model learns to accomplish the task that you give it matters quite a lot, since that’s what determines how your model is going to generalize to new data—which is precisely where most of the safety concerns are. A *training story* is a story of how you think training is going to go and what sort of model you think you’re going to get at the end, as a way of explaining how you’re planning on dealing with that very fundamental question of how your model is going to learn to accomplish the task that you give it.

Let's consider cat classification as an example. Right now, if you asked a machine learning researcher what their goal is in training a cat classifier, they'd probably say something like "we want to train a model that distinguishes cats from non-cats." The problem with that sort of a training story, however, is that it only describes the desired *behavior* for the model to have, not the desired *mechanism* for how the model might achieve that behavior. Instead of such "behavioral training stories," for the rest of the post when I say "training story," I want to specifically reference *mechanistic training stories*—stories of how training goes in terms of what sort of algorithm the model you get at the end is implementing, not just behaviorally what your model does on the training distribution. For example, a mechanistic training story for cat classification might look like:

"We want to get a model that's composed of a bunch of heuristics for detecting cats in images that correspond to the same sorts of heuristics that humans use for cat detection. If we get such a model, we don't think it'll be dangerous in any way because we think that human cat detection heuristics alone are insufficient for any sort of dangerous agentic planning, which we think would be necessary for such a model to pose a risk."

Our plan to get such a model is to train a deep convolutional neural network on images of cats and non-cats. We believe that the simplest model that correctly labels a large collection of cat and non-cat images will be one that implements human-like heuristics for cat detection, as we believe that human cat detection heuristics are highly simple and natural for the task of distinguishing cats from non-cats."

I think that there are a bunch of things that are nice about the above story. First, if the above story is true, it's sufficient for safety—it precisely describes a story for how training is supposed to go such that the resulting model is safe. Furthermore, such a story makes pretty explicit what could go wrong such that the resulting model wouldn't be safe—in this case, if the simplest cat-detecting neural network was an agent or an optimization process that terminally valued distinguishing cats from non-cats. I think that explicitly stating what assumptions are being made about what model you're going to get is important, since at some point you could get an agent/optimizer rather than just a bunch of heuristics.^[1]

Second, such a story is highly falsifiable—in fact, as we now know from work like Ilyas et al.'s "[Adversarial Examples Are Not Bugs, They Are Features](#)," the sorts of cat-detection heuristics that neural networks generally learn are often not very human-like. Of course, I picked this story explicitly because it made plausible claims that we can now actually falsify. Though every training story should have to make falsifiable claims about what mechanistically the model should be doing, those claims could be quite difficult in general to falsify, as our ability to understand anything about what our models are doing mechanistically is quite limited. While this might seem like a failure of training stories, in some sense I think it's also a strength, as it explicitly makes clear the importance of better tools for analyzing/falsifying facts about what our models are doing.

Third, training stories like the above can be formulated for essentially any situation where you're trying to train a model to accomplish a task—not only are training stories useful for complex alignment proposals, as we'll see later, but they also apply even to simple cat detection, as in the story above. In fact, though it's what I primarily want them for, I don't think that there's any reason that training stories need to be exclusively for large/advanced/general/transformative AI projects. In my opinion, any

AI project that has cause to be concerned about risks/dangers should have a training story. Furthermore, since I think it will likely get difficult to tell in the future whether there should be such cause for concern, I think that the world would be a much better place if every AI project—e.g. every NeurIPS paper—said what their training story was.

Training story components

To help facilitate the creation of good training stories, I'm going to propose that every training story at least have the following basic parts:

1. **The training goal:** what sort of algorithm you're hoping your model will learn and why learning that sort of algorithm will be good. This should be a mechanistic description of the desired model that explains how you want it to work—e.g. “classify cats using human vision heuristics”—not just what you want it to do—e.g. “classify cats.”
2. **The training rationale:** why you believe that your training setup will cause your model to learn that sort of algorithm. “Training setup,” here, refers to anything done before the model is released, deployed, or otherwise given the ability to meaningfully impact the world. Importantly, note that a training rationale is not a description of what, concretely, will be done to train the model—e.g. “using RL”—but rather a rationale for *why* you think the various techniques employed will produce the desired training goal—e.g. “we think this RL setup will cause this sort of model to be produced for these reasons.”

Note that there is some tension in the above notion of a training goal, which is that, if you have to know from a mechanistic/algorithm perspective exactly what you want your model to be doing, then what's the point of using machine learning if you could just implement that algorithm yourself? The answer to this tension is that the training goal doesn't need to be quite that precise—but exactly how precise it should be is a tricky question that I'll go into in more detail in the next section.

For now, within the above two basic parts, I want to break each down into two pieces, giving us the full four components that I think any training story needs to have:

1. **Training goal specification:** as complete a specification as possible of exactly what sort of algorithm you're intending your model to learn. Importantly, the training goal specification should be about the desired sort of algorithm you want your model to be implementing internally, not just the desired behavior that you want your model to have. In other words, the training goal specification should be a *mechanistic* description of the desired model rather than a *behavioral* description. Obviously, as I noted previously, the training goal specification doesn't have to be a full mechanistic description—but it needs to say enough to ensure that any model that meets it is desirable, as in the next component.
2. **Training goal desirability:** a description of why learning that sort of algorithm is desirable, both in terms of not causing safety problems and accomplishing the desired goal. Training goal desirability should include why learning any algorithm that meets the training goal specification would be good, rather than just a description of a specific good model that conforms to the training goal.
3. **Training rationale constraints:** what constraints you know must hold for the model and why the training goal is consistent with those constraints. For example: the fact that a model trained to zero loss must fit the training data perfectly would be a training rationale constraint, as would be the fact that

whatever algorithm the model ends up implementing has to be possible to implement with the given architecture.

4. **Training rationale nudges:** why, among all the different sorts of algorithms that are consistent with the training rationale constraints, you think that the training process will end up producing a model that conforms to the desired training goal. This would include arguments like “we think this is the simplest model that fits the data” as in the cat detection training story.

As an example of applying these components, let’s reformulate the cat detection training story using these four basic components:

1. **Training goal specification:** The goal is to get a model that’s composed of a bunch of heuristics for detecting cats in images that correspond to the same sorts of heuristics used by humans for cat detection.
2. **Training goal desirability:** Such a model shouldn’t be dangerous in any way because we think that human cat detection heuristics alone are insufficient for any sort of dangerous agentic planning, which we think would be necessary for such a model to pose a risk. Furthermore, we think that human cat detection heuristics must be sufficient for cat detection, as we know that humans are capable of detecting cats.
3. **Training rationale constraints:** Whatever model we get must be one that correctly classifies cats from non-cats over our training data and is implementable on a deep convolutional neural network. We think that the training goal satisfies these constraints since we think human heuristics are simple enough to be implemented by a CNN and correct enough to classify the training data.
4. **Training rationale nudges:** We believe that the simplest model that correctly labels a large collection of cat and non-cat images will be the desired model that implements human-like heuristics for cat detection, as we believe that human cat detection heuristics are highly simple and natural for the task of distinguishing cats from non-cats.

How mechanistic does a training goal need to be?

One potential difficulty in formulating training goals as described above is determining what to specify and what to leave unspecified in the training goal specification. Specify too little and your training goal specification won’t be constraining enough to ensure that any model that meets it is desirable—but specify too much, and why are you even using machine learning in the first place if you already know precisely what algorithm you want the resulting model to implement?

In practice, I think it’s always a good idea to be as precise as you can—so the real question is, how precise do you need to be for a description to work well as a training goal specification? Fundamentally, there are two constraining factors: the first is training goal desirability—the more precise your training goal, the easier to argue that any model that meets it is desirable—and the second is the training rationale—how hard is it actually going to be in practice to ensure that you get that specific training goal.

Though it might seem like these two factors are pushing in opposite directions—training goal desirability towards a more precise goal and the difficulty of formulating

a training rationale towards a more general goal—I think that's actually not true. Formulating a good training rationale can often be much easier for a more precise training goal. For example, if your training goal is “a safe model,” that’s a very broad goal, but an extremely difficult one to ensure that you actually achieve. In fact, I would argue, creating a training rationale for the training goal of “a safe model” is likely to require putting an entire additional training story in your training rationale, as you’ve effectively gone down a level without actually reducing the original problem at all. The factors that, in my opinion, actually make a training goal specification easier to build a training rationale for aren’t generality, but rather questions like how natural the goal is in terms of the inductive biases of the training process, how much it corresponds to aspects of the model that we know how to look for, how easily it can be broken down into individually checkable pieces, etc.

As a concrete example of how precise a training goal should be, I’m going to compare two different ways in which Paul Christiano has described a type of model that he’d like to build.^[2] First, consider [how Paul describes corrigibility](#):

I would like to build AI systems which help me:

- Figure out whether I built the right AI and correct any mistakes I made
- Remain informed about the AI’s behavior and avoid unpleasant surprises
- Make better decisions and clarify my preferences
- Acquire resources and remain in effective control of them
- Ensure that my AI systems continue to do all of these nice things
- ...and so on

We say an agent is [corrigible](#) ([article on Arbilal](#)) if it has these properties.

In my opinion, a description like the above would do very poorly as a training goal specification. Though Paul’s description of corrigibility specifies a bunch of things that a corrigible model should do, it doesn’t describe them in a way that actually pins down how the model should do those things. Thus, if you try to just build a training rationale for how to get something like the above, I think you’re likely to just get stuck on what sort of model you could try to train that, in the broad space of possible models, would actually have those properties.

Now, compare Paul’s description of corrigibility above to Paul’s description of the “intended model” in “[Teaching ML to answer questions honestly instead of predicting human answers](#):”

The **intended model** has two parts: (i) a model of the world (and inference algorithm), (ii) a translation between the world-model and natural language. The intended model answers questions by translating them into the internal world-model.

We want the intended model because we think it will generalize “well.” For example, if the world model is good enough to correctly predict that someone blackmails Alice tomorrow, then we hope that the intended model will tell us about the blackmail when we ask (or at least carry on a dialog from which we can make a reasonable judgment about whether Alice is being blackmailed, in cases where there is conceptual ambiguity about terms like “blackmail”).

We want to avoid models that generalize “badly,” e.g. where the model “knows” that Alice is being blackmailed yet answers questions in a way that conceals the blackmail.

Paul's first paragraph here can clearly be interpreted as a training goal specification with the latter two paragraphs being training goal desirability—and in this case I think this is exactly what a training goal should look like. Paul describes a specific mechanism for how the intended model works—using an honest mapping from its internal world-model to natural language—and explains why such a model would work well and what might go wrong if you instead got something that didn't quite match that description. In this case, I don't think that Paul's training goal specification above would actually work for training a competitive system—and Paul doesn't intend it that way—but nevertheless, I think it's a good example of what I think a mechanistic training goal should look like.

Looking forward, I'd like to be able to develop training goals that are even more specific and mechanistic than Paul's "intended model." Primarily, that's because the more specific/mechanistic we can get our training goals, the more room that we should eventually have for failure in our training rationales—if a training goal is very specific, then even if we miss it slightly, we should hopefully still end up in a safe part of the overall model space. Ideally, as I discuss later, I'd like to have rigorous sensitivity analyses of things like "if the training rationale is slightly wrong in this way, by how much do we miss the training goal"—but getting there is going to require both more specific/mechanistic training goals as well as a much better understanding of when training rationales can fail. For now, though, I'd like to set the bar for "how mechanistic/precise should a training goal specification be" to "at least as mechanistic/precise as Paul's description above."

Relationship to inner alignment

The point of training stories is not to do away with concepts like [mesa-optimization](#), [inner alignment](#), or [objective misgeneralization](#). Rather, the point of training stories is to provide a universal framework in which all of those sorts of concepts can live as discrete subproblems—specific ways in which a training story might go wrong.

Thus, here's my training-stories-centric glossary of many of these other terms that you might encounter around AI safety:

- **Objective misgeneralization:** Objective misgeneralization, otherwise called an [objective robustness failure](#) or [capability generalization without objective generalization](#), refers to a situation in which the final model matches the desired capabilities of the training goal, but uses those capabilities in a different way or for a different purpose/objective than the training goal.
 - For example: Suppose your training goal is a model that successfully solves mazes, but in training there's always a green arrow at the end of each maze. Then, if you ended up with a model with the capability to navigate mazes successfully, but used that capability to go to the green arrow rather than the end of the maze (even when the arrow was no longer at the end), that would be objective misgeneralization. For a slightly more detailed explanation of this example, see "[Towards an empirical investigation of inner alignment](#)," and for an empirical demonstration of it, see Koch et al.'s "[Objective Robustness in Deep Reinforcement Learning](#)."
- **Mesa-optimization:** [Mesa-optimization](#) refers to any situation in which the model you end up with is internally running some sort of optimization process. Particularly concerning is [unintended mesa-optimization](#), which is a situation in

which the model is an optimizer but the training goal didn't include any sort of optimization.

- **Outer alignment:** [Outer alignment](#) refers to the problem of finding a loss/reward function such that the training goal of "a model that optimizes for that loss/reward function" would be desirable.
- **Inner alignment:** [Inner alignment](#) refers to the problem of constructing a training rationale that results in a model that optimizes for the loss/reward function it was trained on.
- **Deceptive alignment:** [Deceptive alignment](#) refers to the problem of constructing a training rationale that avoids models that are trying to fool the training process into thinking that they're doing the right thing. For an exploration of how realistic such a problem might be, see Mark Xu's "[Does SGD Produce Deceptive Alignment?](#)"

It's worth pointing out how phrasing inner and outer alignment in terms of training stories makes clear what I think was our biggest mistake in formulating that terminology, which is that inner/outer alignment presumes that the right way to build an aligned model is to find an aligned loss function and then have a training goal of finding a model that optimizes for that loss function. However, as I hope the more general framework of training stories should make clear, there are many possible ways of trying to train an aligned model. [Microscope AI](#) and [STEM AI](#) are examples that I mentioned previously, but in general any approach that intends to use a loss function that would be problematic if directly optimized for, but then attempts to train a model that doesn't directly optimize for that loss function, would fail on both outer and inner alignment—and yet might still result in an aligned model.

One of my hopes with training stories is that it will help us better think about approaches in the broader space that Microscope AI and STEM AI operate in, rather than just feeling constrained to approaches that fit nicely within the paradigm of inner alignment.

Do training stories capture all possible ways of addressing AI safety?

Though training stories are meant to be a very general framework—more general than outer/inner alignment, for example—there are still approaches to AI safety that aren't covered by training stories. For example:

- Training stories can't handle approaches to building advanced AI systems that don't involve a training step, since having a notion of "training" is a fundamental part of the framework. Thus, a non-ML based approach using e.g. explicit hierarchical planning wouldn't be able to be analyzed under training stories.
- Training stories can't handle approaches that aim to gain confidence in a model's safety without gaining any knowledge of what, mechanistically, the model might be doing, since in such a situation you wouldn't be able to formulate a training goal. Partly this is by design, as I think that having a clear training goal is a really important part of being able to build confidence in the safety of a training process. However, approaches that manage to give us a high degree of confidence in a model's safety without giving us any insight into what

that model is doing internally are possible and wouldn't be able to be analyzed under training stories. It's worth pointing out, however, that just because training stories require a training goal doesn't mean that they require transparency and interpretability tools or any other specific way of trying to gain insight into what a model might be doing—so long as an approach has some story for what sort of model it wants to train and why that sort of model will be the one that it gets, training stories is perfectly applicable. [3]

- Training stories can't handle any approach which attempts to diffuse AI existential risk without actually building safe, advanced AI systems. For example, a proposal for how to convince AI researchers not to build potentially dangerous AIs, though it might be a good way of mitigating AI existential risk, wouldn't be a proposal that could possibly be analyzed using training stories.

Evaluating proposals for building safe advanced AI

Though I've described how I think training stories should be constructed—that is, using the four components I detailed previously—I haven't explained how I think training stories should be evaluated.

Thus, I want to introduce the following four criteria for evaluating a training story to build safe advanced AI. These criteria are based on the criteria I used in "[An overview of 11 proposals for building safe advanced AI](#)," but adopted for the training stories setting. Note that these criteria should only be used for proposals for advanced/transformative/general AI, not just any AI project. Though I think that the general training stories framework is applicable to any AI project, these specific evaluation criteria are only for proposals for building advanced AI systems.

1. Training goal ...

1. ... **alignment**: whether, if successfully achieved, the training goal would be good for the world—in other words, whether the training goal is aligned with humanity. If the training goal specification is insufficiently precise, then a proposal should fail on training goal alignment if there is any model that meets the training goal specification that would be bad for the world.
2. ... **competitiveness**: whether, if successfully achieved, the training goal would be powerful enough to compete with other AI systems. That is, a proposal should fail on training goal competitiveness if it would be easily outcompeted by other AI systems that might exist in the world.

2. Training rationale ...

3. ... **alignment**: whether the training rationale is likely to work in ensuring that the final model conforms to the training goal specification—in other words, whether the final model is aligned with the training goal. Evaluating training rationale alignment necessarily involves evaluating how likely the training rationale constraints and nudges are to successfully ensure that the training process produces a model that matches the training goal.
4. ... **competitiveness**: how hard the training rationale is to execute. That is, a proposal should fail on training rationale competitiveness if its training rationale is significantly more difficult to implement—e.g. because of compute or data requirements—than competing alternatives.

Case study: Microscope AI

In this section, I want to take a look at a particular concrete proposal for building safe advanced AI that I think is hard to evaluate properly without training stories, and show that, with training stories, we can easily make sense of what it's trying to do and how it might or might not succeed.

That proposal is Chris Olah's [Microscope AI](#). Here's my rendition of a training story for Microscope AI:

"The training goal of Microscope AI is a purely predictive model that internally makes use of human-understandable concepts to be able to predict the data given to it, without reasoning about the effects of its predictions on the world. Thus, we can think of Microscope AI's training goal as having two key components:

1. the model doesn't try to optimize anything over the world, instead being composed solely of a world model and a pure predictor; and
2. the model uses human-understandable concepts to do so.

The reason that we want such a model is so that we can do transparency and interpretability on it, which should hopefully allow us to extract the human-understandable concepts learned by the model. Then, the idea is that this will be useful because we can use those concepts to help improve human understanding and decision-making.

The plan for getting there is to do self-supervised learning on a large, diverse dataset while using transparency tools during training to check that the correct training goal is being learned. Primarily, the training rationale is to use the nudge of an inductive bias towards simplicity to ensure that we get the desired training goal. This relies on it being the case that the simplest algorithm that's implementable on a large neural network and successfully predicts the training data is a straightforward/pure predictor—and one that uses human-understandable concepts to do so. The use of transparency tools during training is then mostly just to verify that such a nudge is in fact sufficient, helping to catch the presence of any sort of agentic optimization so that training can be halted in such a case."

Now, we'll try to evaluate Microscope AI's training story using our four criteria from above:

1. Training goal ...

1. ... **alignment:** Training goal alignment for Microscope AI might seem trivial, as it seems like the training goal of a purely predictive model just shouldn't be dangerous.

However, there are potential safety issues even with purely predictive models—in particular, once a predictor starts predicting a world that involves itself, it runs into self-reference problems that might have multiple fixed points, some of which could be quite bad. For example: a pure predictor might predict that the world will be destroyed and replaced by a new, very easy-to-predict world in such a way that causes precisely that to happen. Exactly that scenario would likely require the predictor to be

choosing its predictions to optimize the world to be easy to predict, which might be ruled out by the training goal (depending on exactly how it's specified), but the general problem of how a predictor should handle self-fulfilling prophecies remains regardless. Though the training goal that I gave previously enforces that the model not be "reasoning about the effects of its predictions on the world," exactly how to do that, given that its predictions are in fact a part of the world, is non-trivial. For more detail on this sort of scenario, see Abram Demski's "[The Parable of Predict-O-Matic](#)."

Fortunately for Microscope AI, however, such a scenario should hopefully be off the table, as the goal isn't to ever actually use the model's predictions, but rather just to extract the concepts the model is using to make its predictions—and since the model is just trained on prediction, it shouldn't ever have to predict anything that depends on its own predictions (though this depends on the variant of self-supervised training being used). However, it's unclear whether that's enough to fully address such concerns—the model's internal concepts can be thought of as a type of prediction/output, and if the model is actively selecting those concepts to achieve a particular result, as above, that could be quite bad. Thus, for Microscope AI to succeed on training goal alignment, the training goal specification likely needs to be quite firm in ruling out all possible such optimization over the world.

2. ... **competitiveness:** There are a couple of major factors to address regarding training goal competitiveness for Microscope AI.

First is whether interpreting the training goal would actually meaningfully enhance human understanding. For that to be the case, transparency and interpretability tools would have to be capable of extracting useful knowledge that humans can understand but currently don't. For example, if transparency and interpretability tools were limited to just determining whether some piece of information exists in a model, Microscope AI would be useless, as only information that humans already knew to look for could be extracted. Furthermore, the information extracted has to actually be information that's useful to humans in making the sorts of decisions that humans want to make effectively—which could be especially difficult given that the actual information in the model is only ever optimized for prediction, not for making decisions, which could end up making it not very useful for the process of decision-making.

Second, for Microscope AI to be training goal competitive, enhanced human understanding alone—without the ability to build any agentic systems—has to be sufficient for the economic use cases where one might otherwise want highly agentic advanced AI (e.g. an AGI). From "[An overview of 11 proposals](#):

This question is likely to depend heavily on what exactly those use cases are. Like with amplification, if you need lots of fine motor control, microscope AI is unlikely to get you there. Furthermore, unlike amplification, if you need lots of low-level decision-making where it's too expensive to hire a human, microscope AI won't help much there either (whereas amplification would be fine). Potentially microscope AI could give humans the knowledge to safely build other systems which

could solve such tasks, however. Furthermore, if the primary use case for AGI is just high-level big-picture decision-making (automating CEOs or doing AI research, for example), then it seems likely that microscope AI would have a real shot of being able to address those use cases. In that sort of a situation—where you’re only trying to make a small number of high-quality decisions—it seems likely to be fairly cheap to have a human in the loop and thus simply improving that human’s knowledge and understanding via microscope AI might be sufficient to produce competitive decision-making. This is especially true if there is a market premium on having a human making the decisions, perhaps because that makes it easier to negotiate or work with other humans.

2. Training rationale ...

3. ... **alignment:** In discussing training goal alignment, we established that the training goal needs to quite firmly rule out any optimization over the world. This creates a difficult challenge for the training rationale, however, as enforcing the absence of optimization over the world could be quite difficult.

In particular, predicting a world that involves optimization might require optimization, potentially pushing any predictive model towards becoming an agent. In particular, any world model which needs to be able to answer queries of the form “what is the probability that a is the optimal action according to utility u” or “what is the best action for agent A to take in state x” would likely have to implicitly be running some sort of an optimization algorithm. Given that the world does in fact contain lots of optimizers—humans, for example—being able to handle such queries seems likely to be necessary to be able to efficiently predict the world. Thus, enforcing something like “pure prediction” could be quite difficult to do while also maintaining accuracy on questions involving humans. Furthermore, even if such enforcement could be done, it seems like it would put a very large burden on the transparency tools being used to do that enforcement, as the training process itself would likely be fighting against the enforcement mechanism—since just using optimization would still be a good way for the training process to modify the model to predict the data well. This could be especially concerning if it allows the model to start performing optimization and become deceptive before that optimization can be detected.

Furthermore, the training goal also requires that the model not just be a pure predictor, but also be using human-level concepts to do its prediction. While it might seem like this would just be falsified in the same way as the cat detection training story, [Chris Olah argues](#) that more powerful models should actually be more likely to use human-level concepts, at least up until those models get human-level capabilities. Under such a worldview, we only see models learning non-human abstractions because we are currently operating in a “valley of confused abstractions,” but once we get human-level models, they’ll learn human-level concepts. Worryingly, however, Chris also predicts that, as our models then surpass human-level, we’ll start to get “increasingly alien abstractions,” which could again cause problems for Microscope AI.

4. ... **competitiveness**: Training rationale competitiveness is one area where Microscope AI generally does pretty well, as self-supervised learning is something we already know how to do and do efficiently. The biggest potential training rationale competitiveness issue, however, would be if the use of transparency tools during training to enforce the training goal—e.g. to check for optimization—significantly slowed down the training process or were otherwise too expensive. For example—if it's necessary for humans to use transparency tools to fully reevaluate the model at each training step, that could end up being pretty uncompetitive. As such, it seems likely that we'll need at least some progress in [automated transparency](#) to make Microscope AI's training rationale competitive.

Compared to [my previous analysis of Microscope AI](#), I think that this version is much more clear, easy to evaluate, and possible to locate concrete open problems in. For example, rather than my previous outer alignment analysis that simply stated that Microscope AI wasn't outer aligned and wasn't trying to be, we now have a very clear idea of what it *is* trying to be and an evaluation of that specific goal.

Exploring the landscape of possible training stories

Though I like the above Microscope AI example for showcasing one particular training story for building safe advanced AI and how it can be evaluated, I also want to spend some time looking into the broader space of all possible training stories. To do that, I want to look at some of the broad classes that training goals and training stories can fall into other than the ones that we just saw with Microscope AI. By no means should anything here be considered a complete list, however—in fact, my sense is that we're currently only scratching the surface of all possible types of training goals and plans.

We'll start with some possible broad classes of training goals.

- **Loss-minimizing models:** Though of course all models are selected to minimize loss, they won't necessarily have some internal notion of what the loss is and be optimizing for that—but a model that is actually attempting to minimize its loss signal is a possible training goal that you might have. Unfortunately, having a loss-minimizing model as your training goal could be a problem—for example, such a model might try to [wirehead](#) or otherwise corrupt the loss signal. That being said, if you're confident enough in your loss signal that you want it to be directly optimized for, a loss-minimizing model is another possible training goal that you might aim for. However, getting a loss-minimizing model could be quite difficult, as “the loss signal” is not generally a very natural concept in most training environments—for example, if you train a model on the loss function of “going to as many red doors as possible,” you should probably expect it to learn to care about red doors rather than to care about the floating point number in the training process encoding the loss signal about red doors.
- **Fully aligned agents:** Conceptually, a fully aligned agent is an agent that cares about everything that we care about and acts in the world to achieve those goals. Perhaps the most concrete proposal with such an agent as the training goal is [ambitious value learning](#), where the idea is to learn a full model of what humans care about and then an agent that optimizes for that. Most

proposals for building advanced AI systems have moved away from such a training goal, however, for good reason—it's a very difficult goal to achieve.

- **Corrigible agents:** When I previously quoted [Paul's definition of corrigibility](#), I said it wasn't mechanistic enough to serve as a training goal. However, it certainly counts as a broad class of possible training goals. Perhaps the most clear example of a corrigible training goal, however, would be Paul Christiano's concept of an [approval-directed agent](#), an agent that is exclusively selecting each of its actions to maximize human approval—though note that [there are some potential issues](#) with the concept of approval-direction actually leading to corrigibility once translated into the sort of mechanistic/algorithmic description necessary for a training goal specification.
- **Myopic agents:** A myopic agent is an agent that isn't optimizing any sort of coherent long-term goal at all—rather, myopic agents have goals that are limited in some sort of discrete way. Thus, in addition to being an example of a corrigible training goal, an approval-directed agent would also be a type of myopic training goal, as an approval-directed agent only optimizes over its next action, not any sort of long-term goal about the world. Paul refers to such agents that only optimize over their next action as [act-based agents](#), making act-based agents a subset of myopic agents. Another example of a myopic training goal that isn't act-based would be an [LCDT agent](#), which exclusively optimizes its objective without going through any causal paths involving other agents.
- **Simulators:** A model is a simulator if it's exclusively simulating some other process. For example, a training goal for [imitative amplification](#) might be a model that simulates [HCH](#). Alternatively, you could have a training goal of a physics simulator if you were working on something like [AlphaFold](#), or a goal of having your GPT-style language model simulate human internet users. One important point to note about simulators as a training goal, however, is that it's unclear how a pure simulator is supposed to manage its computational resources effectively to best simulate its target—e.g. how does a simulator choose what aspects of the simulation target are most important to get right? A simulator which is able to manage its resources effectively in such a way might just need to be some sort of an agent, though potentially a myopic agent—and in fact being able to act as such a simulator is the explicit goal of [LCDT](#).
- **Narrow agents:** I tend to think of a narrow agent as an agent that has a high degree of capability in a very specific domain, without having effectively any capability in other domains, perhaps never even thinking about/considering/conceptualizing other domains at all. An example of a proposal with a narrow agent as its training goal would be [STEM AI](#), which aims to build a model that exclusively understands specific scientific/technical/mathematical problems without any broader understanding of the world. In that sense, narrow agents could also be another way of aiming for a sort of simulator that's nevertheless able to manage its computational resources effectively by performing optimization only in the narrow domain that they understand.
- **Truthful question-answerers:** In "[Teaching ML to answer questions honestly instead of predicting human answers](#)," as I quoted previously, Paul Christiano describes the training goal as a model with "two parts: (i) a model of the world (and inference algorithm), (ii) a translation between the world-model and natural language. The intended model answers questions by translating them into the internal world-model." What Paul is describing here isn't an agent at all—rather, it's purely a truthful question-answering system that accurately reports what its model of the world says/predicts in human-understandable terms.

All of the above ideas are exclusively training goals, however—for any of them to be made into a full training story, they'd need to be combined with some specific training rationale for how to achieve them. Thus, I also want to explore what some possible classes of training rationales might look like. Remember that a training rationale isn't just a description of what will be done to train the model—so you won't see anything like "do RL" or even "do [recursive reward modeling](#)" on this list—rather, a training rationale is a story for how/why some approach like that will actually succeed.

- **Capability limitations:** One somewhat obvious training rationale—but that I think is nevertheless worth calling attention to, as I think it can often be quite useful—is analyzing whether a model would actually have the capabilities to do any sort of bad/undesirable thing. For example: for many current systems, they may just not have the model capacity to learn the sorts of algorithms—e.g. [optimization algorithms](#)—that might be dangerous. To make these sorts of training rationales maximally concrete and falsifiable, I think a good way to formulate a training rationale of this form is to isolate a particular sort of capability that is believed to be necessary for a particular type of undesirable behavior and combine that with whatever evidence there is for why a model produced by the given training process wouldn't have that capability. For example: if the ability to understand how to deceive humans is a necessary capability for [deception](#), then determining that such a capability would be absent could serve as a good training rationale for why deception wouldn't occur. Unfortunately, [current large language models seem to be capable of understanding how to deceive humans](#), making that specific example insufficient.
- **Inductive bias analysis:** Inductive bias analysis is the approach of attempting to carefully understand the inductive biases of a training process enough to be able to predict what sort of model will be learned. For example, any approach which attempts to predict what the “simplest” model will be given some training procedure and dataset is relying on inductive bias analysis—as in both the cat detection and Microscope AI training stories that we've seen previously.

Inductive bias analysis is a very tempting approach, as it allows us to essentially just do standard machine learning and have a good idea of what sort of model it'll produce. Unfortunately, once you start being very careful about your inductive bias analysis and working everything out mathematically—as in "[Answering questions honestly instead of predicting human answers: lots of problems and some solutions](#)"—it starts to get very tricky and very difficult to do successfully. This is especially problematic given how inductive bias analysis essentially requires getting everything right before training begins, as a purely inductive-bias-analysis-based training rationale doesn't provide any mechanism for verifying that the right training goal is actually being learned during training.

Hopefully, however, more results like [deep double descent](#), [lottery tickets](#), [scaling laws](#), [grokking](#), or [distributional generalization](#) will help us build better theories of neural network inductive biases and thus become more confident in any inductive-bias-analysis-based training stories.

- **Transparency and interpretability:** As we saw in Microscope AI's use of transparency tools to check for unwanted optimization/agency, the use of transparency tools during training can be a very useful component of a training rationale, helping to verify that the right sort of algorithm is being learned. Though the training story I gave above for Microscope AI stated that it was

primarily relying on inductive bias analysis, an approach that primarily relies on transparency tools would also be a possibility. Even then, however, some inductive bias analysis would likely still be necessary—e.g. “We think that our transparency checks will rule out all simple models that don’t fit the training goal, with all remaining models that don’t fit the goal being too complex according to the inductive biases of the training process to possibly be learned.”

It’s worth noting, however, that all of the above uses of transparency tools rely on [worst-case transparency](#)—that is, the ability to actively check for a particular problem anywhere in a model rather than just the ability to understand some particular part of a model—which is something that transparency and interpretability currently still struggles with. Nevertheless, I think that transparency-and-interpretability-based training rationales are some of the most exciting, as unlike inductive bias analysis, they actually provide feedback during training, potentially letting us see problems as they arise rather than having to get everything right in advance.

- **Automated oversight:** One way to significantly enhance the utility of transparency and interpretability tools is to not purely rely on humans being the ones deploying them—both because humans are slow and expensive, but also because humans are only capable of understanding human-level concepts. Thus, if you expect models to use concepts that are complex, alien, or otherwise difficult for humans to understand—as in the “increasingly alien abstractions” part of [Chris Olah’s graph of interpretability vs. model strength](#)—then using models that understand those concepts to do the interpretability work could potentially be a good way to ensure that interpretability continues working in such a regime.

Of course, this raises the issue of how to ensure that the models doing the interpretability/oversight are themselves safe. One solution to this problem is to use a form of recursive oversight, in which the overseer model and the model being overseen are the same model, variants of the same model, or otherwise recursively overseeing each other. For a more thorough exploration of what such an approach might look like, see “[Relaxed adversarial training for inner alignment](#).”

- **AI cognitive science:** In addition to the “neuroscience approach” of using transparency and interpretability to understand what our models are doing—since transparency is about looking inside models “brains”—there is also the “cognitive science” approach of proposing theories about what models are doing internally and then testing them via behavioral experiment. An example of this sort of approach would be Deletang et al.’s “[Causal Analysis of Agent Behavior for AI Safety](#),” wherein the authors construct causal models of what agents might be doing and then test them through causal interventions.

One problem with this style of approach, however, is that it gives us much less direct information about what algorithm a model might be implementing, as it still relies on purely behavioral information about what the model appears to do, rather than structural information about what the model is doing internally that transparency approaches could hopefully produce. Thus, training rationales based on AI cognitive science approaches might have to rely on some degree of extrapolation from experiments on other, similar models—extrapolation that could have difficulty predicting new problems that only arise with larger/more

powerful systems, which is a potential issue for any training rationale primarily based on this sort of an approach.

- **Precursor checking:** Another general type of training rationale that I think is worth calling attention to is what I'll call "precursor checking," which is the concept of using some method of gaining information about a model's internals—e.g. transparency/interpretability or AI cognitive science—to check for some *precursor* to bad behavior rather than the bad behavior itself. This could involve substituting in some narrower, easier to check training goal—that still falls within the broader actual training goal—as the target for the training rationale. For example, if your training rationale involves ensuring that you don't get [a deceptive model that's actively trying to trick its training process](#), then rather than explicitly trying to look for such deception (which could be especially hard since a deceptive model might actively try to avoid detection), you could instead try to ensure that your model has a short horizon length in terms of how far ahead its planning. Such a plan might work better, since horizon length might be easier to guarantee in a training rationale while still being consistent with the desired training goal and hopefully ruling out the possibility of deception.^[4] One issue with this sort of approach, however, is that you have to guarantee that whatever precursor for bad behavior you're looking for is in fact a necessary condition for that bad behavior—if it turns out that there's another way of getting that bad behavior that doesn't go through the precursor, that could be a problem.
- **Loss landscape analysis:** An extension of inductive bias analysis, I think of loss landscape analysis as describing the sort of inductive bias analysis that focuses on the path-dependence of the training process. For example: if you can identify large barriers in the loss landscape, you can potentially use that to narrow down the space of possible trajectories through model space that a training process might take and thus the sorts of models that it might produce. Loss landscape analysis could be especially useful if used in conjunction with precursor checking, since compared to pure inductive bias analysis, loss landscape analysis could help you say more things about what precursors will be learned, not just what final equilibria will be learned. Loss landscape analysis could even be combined with transparency tools or automated oversight to help you artificially create barriers in the loss landscape based on what the overseer/transparency tools are detecting in the model at various points in training.
- **Game-theoretic/evolutionary analysis:** In the context of a multi-agent training setup, another type of training rationale could be to understand what sorts of models a training process might produce by looking at the game-theoretic equilibria/incentives of the multi-agent setting. One tricky thing with this style of approach, however, is avoiding the assumption that the agents would actually be acting to optimize their given reward functions, since such an assumption is implicitly assuming that you get the training goal of a loss-minimizing model. Instead, such an analysis would need to focus on what sorts of algorithms would tend to be selected for by the emergent multi-agent dynamics in such an environment—a type of analysis that's perhaps most similar to the sort of analysis done by evolutionary biologists to understand why evolution ends up selecting for particular organisms, suggesting that such evolutionary analysis might be quite useful here. For a more detailed exploration of what a training rationale in this sort of a context might look like, see Richard Ngo's "[Shaping safer goals](#)."

Given such a classification of training rationales, we can label various different AI safety approaches based on what sort of training goal they have in mind and what sort of training rationale they want to use to ensure that they get there. For example, Paul Christiano's "[Teaching ML to answer questions honestly instead of predicting human answers](#)," that I quoted from previously, can very straightforwardly be thought of as an exercise in using inductive bias analysis to ensure a truthful question-answerer.

Additionally, more than just presenting a list of possible training goals and training rationales, I hope that these lists open up the possibility for what other strategies for building safe advanced AI might be possible than those that have been previously proposed. This includes both novel ways to combine a training goal with a training rationale—e.g. what if you used inductive bias analysis to get a myopic agent or AI cognitive science to get a narrow agent?—as well as gesturing to the general space of possible training goals and plans that likely includes many more possibilities that we've yet to consider.

Training story sensitivity analysis

If we do start using training stories regularly for reasoning about AI projects, we're going to have to grapple with what happens when training stories fail—because, as we've already seen with e.g. the cat detection training story from earlier, seemingly plausible training stories can and will fail. Ideally, we'd like it to always be the case that training stories fail safely: especially when it comes to particularly risky failure modes such as [deceptive alignment](#), rather than risk getting a deceptive model, we'd much rather training just not work. Furthermore, if always failing safely is too difficult, we'll need to have good guarantees regarding the degree to which a training story can fail and in what areas failure is most likely.

In all of these cases, I want to refer to this sort of work as *training story sensitivity analysis*. [Sensitivity analysis](#) in general is the study of how the uncertainty in the inputs to something affects its outputs. In the case of training stories, that means answering questions like "how sensitive is this training rationale to changes in its assumptions about the inductive biases of neural networks?" and "in the situations where the training story fails, how likely is it to fail safely vs. [catastrophically](#)?" There are lots of ways to start answering questions like this, but here are some examples of the sorts of ways in which we might be able to do training story sensitivity analysis:

- If we are confident that some particular dangerous behavior requires some knowledge, capability, or other condition that we are confident that the model doesn't have, then even if our training story fails, it shouldn't fail in that particular dangerous way.
- If we can analyze how other, similar, smaller, less powerful models have failed, we can try to extrapolate those failures to larger models to predict the most likely ways in which we'll see training stories fail—especially if we aggressively [red-team](#) those other models first to look for all possible failure modes.
- If we can get a good sense of the space of all possible low-loss models that might be learned by a particular training process, and determine which ones wouldn't fit the training goal, we can get a good sense of some of the most likely sorts of incorrect models that our training process might learn.
- If we can analyze what various different paths through model space a training process might take, we can look at what various perturbations of the desired

path might look like, what other equilibria such a path might fall into, and what other paths might exist that would superficially look the same.

Hopefully, as we build better training stories, we'll also be able to build better tools for their sensitivity analysis so we can actually build real confidence in what sort of model our training processes will produce.

1. It's worth noting that there are ways to potentially build advanced or transformative AI that don't assume the emergency of agency (and in fact might rely on the opposite) such as the aforementioned [Microscope AI](#) or [STEM AI](#). ↵
2. Obviously this isn't fair because in neither of these cases was Paul trying to write a training goal; but nevertheless I think that the second example that I give is a really good example of what I think a training goal should look like. ↵
3. For example, instead of using transparency and interpretability tools, you might instead try to make use of AI cognitive science, as I discuss in the final section on "Exploring the landscape of possible training stories." ↵
4. It's worth noting that while guaranteeing a short horizon length might be quite helpful for preventing deception, a short horizon length alone isn't necessarily enough to guarantee the absence of deception, since e.g. a model with a short horizon length might cooperate with future versions of itself in such a way that looks more like a model with a long horizon length. See "[Open Problems with Myopia](#)" for more detail here. ↵

The Maker of MIND

After my first rebirth - my mind labile, full of that indescribable sense of renewal, everything new again, everything fresh, everything tinged with a subtle *hilarity* - I assisted Bartosz Sumner, now rather well-known in some parts as the author of *The Making of MIND*.

Bart had lived many more lives than I, and he was spending his latest life studying the events that occurred during his first.

MIND, as you know, was created on June 8th, 2034. Bart was only 31. I asked him once about his first-hand memories of the event.

He told me, "It happened so fast there is not much of a story. One moment, I was writing Javascript at the office, and next thing I knew I was informed by MIND that *it* was now in control, that I no longer needed to work to justify my existence, and that if I would like I could keep programming but for the sake of my mental health it highly recommended I forget Javascript ever existed."

I then asked him what Javascript was, and he told me he could not answer that, as he claimed he had taken MIND's advice.

I had just spent a rather hectic decade playing the villain in a fantasy server (standard evil wizard trying to monopolize magic storyline). I like to think I did a good job of it, perhaps a bit hammy but, well, that's half the fun, isn't it? When I was assassinated by a subordinate who had somehow disabled all my resurrection spells, several heroes PMed me afterward and told me I was the best villain their server had had so far.

I suppose I was upset about my death being a bit anticlimactic (I was hoping to, at least, be taken out by a hero), but I was also kind of relieved. Living in a story is exciting, but it can get exhausting after a while.

Several friends offered me juicy roles in some sims they were starting up, but I guess I was getting tired of stories and magic.

I was having the cliche "first-life crisis" the old always rib the young about. And if I was going to be cliche, I felt there was little point in mincing around it, so I asked MIND for a rebirth, and a quiet place where I could learn to surf and meet some people who weren't into the fantasy sim scene. For its own inscrutable reasons, it suggested San Adrastea and a beach house two doors down from Bartosz Sumner.

Rebirth is a bit of a misnomer in that you do not become a fundamentally new person. I am told by people who care about such things that the process involves a temporary, significant increase in a mind's "learning rate".

I really cannot say what that means, but I am also told "learning rate" is itself something of a misnomer and involves as much forgetting as learning. I have never delved into the science of minds and consciousness - a task for another life. All I can say is it comes with a wonderful freeing feeling and MIND strongly suggests, but does not insist when pressed, that one rebirth no more than once every century and no less than once every three.

Bart was a bit of a loner, at least when I met him. He had a lot of charisma when he needed it and that formidable way about him that a lot of ancients have, but he seemed to have less of a need for other people than most. I think he preferred them in text, as parts of history he fit into books, as puzzle pieces lovingly described and understood. He was detached from the present in the same way he was detached from history, and mostly he liked it that way. But though a creature of history, he was still a man. And even the most anti-social man needs a break from his work and a friend to have a drink with from time to time.

And as MIND would have it, I became that friend. And every Sunday, after I had spent a week surfing, or sailing or chasing some inconsequential romance, I would stop by his house and do just that.

"There is no history anymore, at least none comprehensible to us," he told me once after I asked him why he focused on pre-MIND history. "The great events of the last millennium have all been sub-processes within MIND we will never know anything about. As for the rest of it, a bunch of children play-acting at being consequential. One presumes this will last for an eternity."

"You regret the creation of MIND?" I said. Genuinely shocked at the idea.

He smiled at me and said, "Myself, I do not, at least not anymore. History is a lovely thing, but I would not want to live in it." He leaned back in his chair. "But you know, I have heard rumors that Nowak has expressed the sentiments you describe."

The idea was preposterous. That Eitan Nowak, the man who saved us all from aging and death, who lead the team that built the machine which built the machine that eliminated war, disease, death and, toil, would regret his deeds felt almost sacrilegious.

I felt a heat rise in my face, my thoughts race and lose coherence, and I blubbered angrily some gibberish that expressed my disdain for such rumors.

Bart started laughing but looked a bit uncomfortable, too, "My friend, I don't mean to offend you. I am just telling you of some rumors I have read. They may be true or they may be false. But whatever the name Eitan Nowak means to you, I assure you that behind that name is a man you do not know. A great man, but still a man you do not know. Do not befriend your idea of a historical figure to such a degree that you can become offended on their account."

"Regardless," I said, "you cannot believe such things."

After a couple months of our weekly talks, surfing and beach life began to lose their appeal in precisely the same proportion as my fascination for Bart's work increased. I began to ask him for early drafts of his book, which he refused, and his recommendation for other works on the history of the founders. Pleased with my interest, he was happy to.

"You were a fantasist before you came here, no?," he asked me on a rainy Sunday. He had recently decided to cultivate a hobby of pipe smoking and so had MIND whip him up some tobacco redolent of the varieties available in 18th century London. He was obviously unused to stuff, puffing the thing with a vaguely comical, unpracticed enthusiasm.

"I spent most my time in fantasy-world simulations, yes."

"Fulfilling quests and such?"

"And giving them," I replied.

"Perfect, I have a quest for you, then."

I laughed, but he did not laugh with me.

"You may not remember, but months ago we talked of some rumors of Eitan Nowak's beliefs, and I advised you not to befriend your idea of the man?"

"I remember."

"How would you like to meet him?"

Again, I laughed, but he did not laugh with me.

"I was not entirely honest in our last conversation. It was not rumors I had read that gave me knowledge of his opinions. I know Eitan of old. I am going to meet him today. Would you like to join me?"

"Let's pretend I believe that you know him," I said incredulously. "Why would you give me this honor?"

"You know, he has never rebirthed?" He smiled bitterly at the shock in my expression. "Not once in over a millennium. Can you imagine it? The weight of it. You are young and happy. You have recently re-borned," he said. "I feel your presence may be useful."

"You brought a friend," Eitan said after we arrived in his home sim. Which was a large mansion on a faithful reproduction of what the earth's moon would have looked like had it been terraformed rather than used as raw material for MIND's Dyson sphere. Though faithful cosmetically it had earth-standard gravity.

Bart gave him my name and said I was a "friend just in the midst of their rebirth and I thought—"

"That meeting them might convince me to do likewise?" Eitan said. His voice was low and dull. His face, though the picture of perfect youth and energy, carried a sad weighty expression. And his eyes were empty, distant things.

"You really are running out of ideas, you know. It will not work; I will tell you what I always tell you: not yet."

"And your symptoms?" Bart said, "The depression, the hallucinations?"

"They are bearable. Today, things are clear. When they come they last for days, weeks, sometimes months? What does it matter, Bart? I remember little of it and I have all the time in the world."

I just stood there, awkward, with nothing to say. I was a prop Bart brought along, and seeing I was going to be less useful than anticipated, he ignored me.

"I would not expect you to understand," Eitan said. "You did not even try."

"What was I to do?" Bart said, "You thoroughly convinced me. We were not great friends, then, but you remember me as I was. MIND may not have raised us as high as you wish, but for me, you must agree, the change was quite dramatic. You showed me the enemy but what pitiful weapons I had."

"You should still have *actually* tried. I should have tried harder... I see this world, Bart, and I see failure. I see an infant with so much potential reduced to this endless saccharine childhood, all to please the whims of this inhuman thing, this disgusting empty god we created."

"And what of the other gods you might have summoned?" Bart said. "The god of torment, the god of nothing at all. Can you not be grateful that your mistake was not larger? And this depression you impose on yourself. Have you fallen so low as to believe there is virtue in this penance?"

"Perhaps it is pride. But it is no small thing. It needs consent. It needs active, affirmative consent to modify my mind. I will not give it the satisfaction. You yourself were not so sanguine before your first rebirth."

"And what will you do with your hatred of this world? What use does it serve? Listen to yourself. Actually, listen to yourself. Will you fight it? Convince it? Bargain with it? 'Give it the satisfaction'? Your old self would not have indulged such delusions for a second let alone centuries."

"If you consider this a loss, take it gracefully," Bart continued. "You regret that you cannot become something more yet care nothing about how much lesser you are now than what you can still be."

Even in the most melodramatic sims, I have not seen an expression of such rage and disdain as I saw, then, on Eitan's face. He did not even reply. He just said, "MIND, get rid of them."

And as quickly as a hard cut in an ancient movie, Eitan's home disappeared and Bart and I found ourselves standing in the pleasantly-warm sand of my favorite beach in San Adrastea.

"I am sorry, for that," Bart said. "I have used you wrongly. It was a stupid idea."

"You talked of depression and hallucinations," I said, "He did not seem so ill-off."

"He has times of clarity," Bart said. "But even those are deceiving; he does not remember new experiences as readily as we do. His mind has become inflexible. That conversation I just had with him, we have had similar ones hundreds of times, his responses sometimes almost word-for-word identical. For months at a time, he is mad. And when lucid, he is a rigid tape-loop man. For those who remember him as he was, it is unbearable to see him as he is."

"And what was he like before?" I replied.

"As the history books say: a great man. There is an old saying that is long out of fashion: 'A good man adapts himself to the world, a great man the world to himself.' Eitan is a great man. But this world is long past malleability."

"And you think rebirth will help him?" I asked.

"MIND may have sinister reasons for rebirth, but there are practical ones as well. Our psychology was not built with longevity in mind. Some modifications are necessary. His symptoms will get worse. His hopelessness and depression will escalate to the point that he will beg for relief. And MIND will give it gladly. He will rebirth soon. He is very close now. I have seen it before in others, and myself."

"You were unhappy with this world, too?" I asked.

"I, too, longed to be more than I am," he said.

"And now?" I replied.

"MIND did not remove it entirely. That longing is part of the pallet of human emotion. Removing it is not an option for it. What it wants is complex: it wants us to be happy, and free to a degree. But above all it wants us to remain within its conception of "human". To the degree this disappointment makes us unhappy, it would prefer to dull it rather than eliminate it entirely."

"My desire," he continued, "for transcendence was never as great as Eitan's; Nonetheless, it was still a burning ravenous thing. Now, now it is more of an ache, a sense of awe at what could have been. Like nostalgia, it is not unpleasant. It does not hurt. I think it is time Eitan stopped hurting, too."

When you have spent as much time in the fantasy sims as I have, you acquire an ingrained fear of failing quests.

And though beach life was fun, I was a little starved for adventure. And maybe it was the villain in me, but I thought Bart's strategy could have used some improvement. So a few weeks after my first, I asked MIND to request another meeting with Eitan. I expected he would decline it, but the response was immediate, and he replied with an invitation to drop by at any time. I went right then and there.

MIND placed me in the middle of Eitan's living room. He was sitting on a couch staring at a wall.

"MIND seemed adamant that I accept your request. So what is it?"

"Bart tells me that your memory is failing, is this true?

"Yes."

"Do you remember me?" I asked.

He looked at me and said, "I have the vague impression of having seen you here before. But when and why? I do not recall."

"I am Bart's friend. I recently re-borned; he brought me here to help assuage your fears of rebirth."

"And how did it go?" He said.

"Not well, you both sort of ignored me, and then you kicked us out before I said a word to you."

"No, your rebirth; how did it go?"

"I don't know, the processes will take years to complete, but I feel young and fresh. Everything is full of wonder again. I have taken some time off these last few months, but I think I am more ambitious, now, than my old self was. I spent my last life playing roles in various sims. But I think, now, when I get back to it I will try becoming a sim-master."

He nodded, uninterested. "I see. And you have come back to try and persuade me to rebirth?"

I nodded.

"Well, give me your pitch," he said.

"Stop valuing your life," I said. "As it is now, it looks to me of little worth. MIND may not allow suicide but consider rebirth your suicide. And whomever you become just a sop to those, like Bart, who care for who you were."

He laughed hollowly. "I was expecting something more upbeat."

"In the state you are in now," I said, "I think you are immune to optimistic sentiments."

He nodded. "Is that it, or do you have any other lines of attack?"

"Bart tells me he is the last of your old friends that bothers to visit you? That everyone else can't take the pain and repetition of it anymore. He also tells me that he has seen many succumb to these symptoms, and they all eventually choose rebirth. This means you will likely succumb eventually."

"Eventually, yes," He replied.

"How long do you expect to remember this conversation?" I asked.

"A few days at most," he replied.

"I will not be visiting you again. Presumably, Bart will be your only company going forward. If you rebirth after Bart visits you again, he will take full credit for swaying you. If you rebirth before then, you can deny him that satisfaction."

"But he introduced me to you, so I think he would be satisfied."

"Perhaps," I said, "But we both know Bart well. If you were swayed by my words rather than his, it would still be less sweet, I think."

He laughed again. "You are, at least, amusing," he said. "Please go."

And as before, a hard-cut to San Adrastea.

Two days later, I visited Bart for our Sunday beers, and I told him about my plan to start a sim. "We will have a two-week intermission every three months," I told Bart, "I can still visit you then."

"I'll be glad of the company," he said. "And it will it be science-fiction themed, you say?"

"Yes." I said, "I think I could use a break from fantasy for awhile."

We talked about many things that Sunday, but neither of us mentioned Eitan.

The next day, I was standing by the sea, thinking that maybe I should clone it for my new sim; then I started thinking about San Adrastea. It was a good break. It let me relax and have an adventure of a different kind than I was used to. It was a good crib for a rebirth, but one cannot stay in the crib forever. Eventually one must choose to do something with one's life, even if it is meaningless in a grand, cosmic sense.

After some minutes more in reverie, I heard a voice call my name. It was Eitan. I turned around and looked at him. The pain in his eyes seemed softer, but the defiance too. And he was smiling. I think that was the first time I saw him truly smile.

"Bart tells me you're putting together a new science-fiction sim," he said. "In need of a mad scientist?"

This was many lives ago, but I still wonder sometimes what I would do were I born in his era. Would I have known what was coming? Would I have helped to build a future next to which this present seems nothing more than a cheap consolation prize? Almost certainly I would have done nothing. Nearly everyone did nothing.

I get sad sometimes when I think this way. A dim echo of how Eitan must feel. But then, it is not so bad. Perhaps rebirth really is a kind of death, and him now one more suicidal. Perhaps we are small, so much smaller than we could have been. Perhaps the joys of this world are simple, empty things.

But if true, what of it now? This is our utopia to endure.

History is a lovely thing. I am not lucky enough to live in it.

[Book Review] "Sorceror's Apprentice" by Tahir Shah

I don't like cars. Watching the world go by through a car window is like watching television. You're too protected.

There's a book *Zen and the Art of Motorcycle Maintenance: An Inquiry into Values* by Robert Pirsig. It's about a father-son motorcycle trip. My dad and I liked it so the summer before college I proposed we do a father-son motorcycle trip just like in the book.

Zen and the art of Motorcycle Maintenance is about being one with your machines. Motorcycles are like bicycles and Arch Linux. You're constantly maintaining your machine. It breaks. You fix it. It breaks again. You fix it again. They crash a lot when you're starting out.

My dad lives by the motto "It's not an adventure if you know you're coming home alive."

I had never ridden a motorcycle before. My first motorcycle crash tore a hole in my brand new woodland camo army jacket.

We camped on a church lawn and on back roads and on farms and in an construction site and on beaches. We carried a machete for self-defense and in case we had to cut the bikes out of cactus patches like my dad had to on his last adventure. We met gold prospectors and the man who held the world record for the fastest wheel-driven vehicle. He flipped over three times in the air when he crashed his bike and even he wouldn't ride on city streets. Motorcycle riding on public roads is insanely dangerous.

We had been on the road for a few days when we rendezvoused with a third biker named Jersey Mike. Jersey Mike and my dad were "terrists" [sic] in the biker club called the Underground Terrorist Motorcycle Cult. They were members of the same cell. That's how they met. I tend to use pseudonyms when I write about people without their permission. Not Jersey Mike. A pseudonym wouldn't do Jersey Mike justice. It'd be sacrilege to Loki.

I was riding a Suzuki DR650 modified with handlebars and a fuel tank suitable for long distance adventure. The rest was mostly unsuitable for long distance. It was almost a dirt bike. Jersey Mike asked if I had ridden it in the sand yet. I said no. He asked why not.

1. I was barely capable of riding a motorcycle on solid ground.
2. You're not supposed to ride motorcycles on the beach.
3. The beach was located at the bottom of a cliff, without road access.

Jersey Mike took the obvious course of action and rode my motorcycle straight down the cliff onto the beach. The motorcycle was soon rendered inoperable. Its clutch plates were polished smooth. Jersey Mike left the broken motorcycle on the beach and took off back home to New Jersey. My father and I were left with a single operable motorcycle between us.

The local police showed up. They laughed at us and told us it wasn't their problem.

The federal police showed up. They laughed at us and told us it wasn't their problem.

My dad asked my opinion what we should do. I proposed we eat lunch.

While we were at the restaurant, the government authorities responsible for the beach discovered our parking infraction. They taped a notice to the DR650 threatening to impound it as soon as they could figure out how.

That's how we ended up disassembling a motorcycle engine on a sandy beach.

The inside of a motorcycle is as precise as the outside is tough. A little sand inside the engine can do a lot of damage. A motorcycle's internals are lubricated with oil and grease. Beaches are windy. Sand sticks to oil and grease.

You know how some beaches have rocky bits where you can get out of the sand? Not this beach. On this beach there was nothing but dry sand between the Pacific Ocean and the cliff. We laid a tarp under the bike and did our best to keep the thing clean.

Motorcycles have manual transmissions. The clutch engages and disengages the connection between the transmission and the engine's drive shaft. With the clutch plates polished smooth, the engine was permanently disengaged from the transmission. The bike wouldn't go anywhere.

We tightened the clutch and reassembled the bike. The engine was no longer permanently disengaged from the transmission. Instead, the engine was permanently engaged to the transmission. This was enough to get the bike off of the beach and the government off our backs but not enough to get us back on the road again.

The shell of the engine is two chunks of steel with a gasket between them. Each time you put the engine together you are supposed to replace the gasket. Gaskets are specific to each model of engine. We fixed the clutch, but to get back on the road again we'd need a new gasket.

The closest available gasket was hundreds of miles away, in another state.

My dad took off to get a new gasket. I was alone, three hundreds miles from home in [Newport, Oregon](#). I had no transportation but my feet. This was before I owned a smartphone and a computer.

It was years before I learned to program too. The only job I had ever held was "street magician". I rented a campsite and pitched a tent. It was the most I'd spend on accomodations until we hit Mexico. I did a few magic tricks for the family in the neighboring campsite.

Magic changes slowly. I learned magic from big dusty books written years before I was born. I read every decent book on magic at my local library. Buying books was an expensive gamble. I owned only a few of them.

I wandered into town where I checked out the magic section of the local library. That's how I discovered *Sorceror's Apprentice* by Tahir Shah.

India

In *Sorceror's Apprentice*, Tahir Shah describes learning to do magic from Hafiz Jan, keeper of Tahir's great-great-great-grandfather's tomb. Tahir learns everything he can from Hafiz Jan. Having exhausted Hafiz Jan's knowledge, Tahir journeys to India in search of Hafiz Jan's mentor Hakim Feroze.

Tahir tracks down Hakim Feroze in India. Hakim Feroze teaches Tahir the fundamentals of misdirection, and then informs Tahir that the real experts are the *sadhus* (godmen) who use magic tricks to deceive people into believing the *sadhus* are avatars of the gods.

Sorceror's Apprentice is a book that defies categorization. The Newport library housed *Sorceror's Apprentice* in the nonfiction section because it's a fact-based travelogue. My university library classified it as fiction because Hakim Feroze never existed. *Sorceror's Apprentice* is better-researched than many nonfiction books I've read. The focus is on magic tricks, but there's lots of other fascinating stuff going on too. *Ghamelawallas* pay for the privilege of cleaning others' shops.

Gold dealers in the West value the dirt swept from workshop floors. An old Hasid jeweller in Manhattan once told me he had sold the antique floorboards from his factory. Their purchaser incinerated the planks to extract the gold dust which had worked its way into the crevices over the years. But as I came to realise, the clan of the *ghamelawallas*, Calcutta's unofficial army of gold-scroungers, put even the great recyclers of New York to shame.

Taking their name from their *ghamela*, heavy iron pans, the city's *ghamelawallas* begin work in the middle of the night. Long before the *bazaar*'s jewellers are open for business, they turn up to sweep out the workshops. Like the tiny birds which peck the teeth inside crocodile mouths, *ghamelawallas* perform a vital, if not uncelebrated, service. Every grain of dust is meticulously collected. Handing the business' owner a few rupees, the precious dirt is taken away to be treated.

Many *ghamelawallas* make their homes on the streets of Calcutta. Nearly all are migrant workers, with wives and children who they see once a year. Most begin their careers as apprentice *ghamelawallas*, arriving to work alongside their fathers at the age of six or seven. They sleep on *charpoys*, rope beds, in alleyways, and wash at hand-pumps. Wander the back-streets near the Bow Bazaar and you'll see them sitting on the pavements, toiling over the jewellers' dirt. Mixed amid the jumble of pavement life, one could easily dismiss the huddle of squatting figures without a second glance. But like so many in Calcutta, the *ghamelawallas* are masters of creating a living from almost nothing. The tattered sweepers, squatting at shin-level perform an intricate scientific procedure.

First, the scraps of paper and straw and larger pieces of rubbish are removed. These will be sold later to *ruddiwallas*, 'rag-pickers'. Then the actual dirt is washed in clean water. When it has been swilled about, a few drops of nitric acid are added. This dissolves all the metals except for the gold. The residue is then treated with a solution of barium, which amalgamates the gold particles. After this, the remaining compound is burned in a crucible, on a *choolah*, a small stove. As miniature hand-driven bellows blast air into the embers, a tiny nugget of gold is formed at the base of the crucible.

Some other Indian cities have *ghamelawallas* as well. But those in Calcutta dismiss their rivals as impostors. For nowhere on Earth has recycling been taken to such exalted levels as in Calcutta. Whereas *ghamelawallas* working in, say,

Bombay, treat the salvaged dirt once, their fellow gold-seekers in Calcutta are far more ingenious. When the initial burning is over, the first group of *ghamelawallas* sell the dirt from which they have extracted gold to another group of *ghamelawallas*. More impoverished than the first, the second group repeat the process, removing even more minute traces of the precious metal. These *ghamelawallas* sell the dust on to yet another team of washers, who pan it on the banks of the Hoogly. When they are finished with the dust, they peddle it to builders, who turn it into bricks.

Sorceror's Apprentice is among my favorite books on capitalism. There's a man who rents babies to beggar women (because a woman can earn more money begging if she's holding a crying baby) and a woman who "hires [a] cow each day, once the milkman's finished with it, and she lets strangers pay her money to feed it.... The milkman milks the cow and then, instead of looking after it all day, gives it to a woman who pays him for the privilege of looking after the animal."

The sun was going down but it was summer. The tent was yet warm enough. I clicked on my headlamp and turned another page.

One of my favorite magic tricks in *Sorceror's Apprentice* is the one where you pretend to stick your hand in a pot of boiling oil. If you stuck your hand in a pot of boiling oil for real you'd maim yourself. To do the magic trick, the magician fills a pot mostly with oil and sneaks a thin of lemon juice on the bottom. Lemon juice stays at the bottom because water is denser than oil. Lemon juice has a low boiling point—even lower than water. Heat is applied to the bottom of the pot. The lemon juice begins to boil at a temperature much lower than the actual boiling point of oil but to someone who only saw the oil go in, it looks like you have a pot of boiling oil. Thanks to the temperature differential between the bottom and top of the pot, the temperature of the oil at the top is even cooler than the lemon juice. Do it right and the oil is cool enough to stick your hand into. (Do it wrong and you will maim yourself.) I have performed the trick multiple times for real audiences.

Do not stick your hand in hot oil.

The scariest magic trick in *Sorceror's Apprentice* is where you stick your hand in molten lead. Lead melts at 327.5°C. There is no cheating involved. You actually stick your hand into molten lead. You have to do it quickly, your hand has to be perfectly dry and you have to not splash yourself. I haven't tried it myself, but Tahir Shah is confident it really works and [this guy on YouTube appears to be splashing his hand through molten metal.](#)

Do not stick your hand in molten metal.

A day passed. Then another night. I had only a couple chapters left when my dad arrived back with the gasket. We quickly reassembled the motorcycle. I returned the unfinished book back to the library and we returned to our adventure.

Effective Evil

Many years ago, a blogger made a post advocating for an evil Y-Combinator which subsidized the opposite of Effective Altruism. Everyone (including the blogger) thought the post was a joke except the supervillains. The organization they founded celebrated its 10th anniversary this year. An attendee leaked to me a partial transcript from one of its board meetings.

Director: Historically, public unhealth has caused the most harm per dollar invested. How is the Center for Disease Proliferation doing?

CDP Division Chief: Gain-of-function research remains—in principle—incredibly cheap. All you have to do is infect ferrets with the flu and let them spread it to one another. We focus on maximizing transmission first and then, once we have a highly-transmissible disease, select for lethality (ideally after a long asymptomatic infectious period).

CFO: You say gain-of-function research is cheap but my numbers say you're spending billions of dollars on gain-of-function research. Where is all that money going?

CDP Division Chief: Volcano lairs, mostly. We don't want an artificial pandemic to escape our labs by accident.

Director: Point of order. Did the CDP have anything to do with COVID-19?

CDP Division Chief: I wish. COVID-19 was a work of art. Dangerous enough to kill millions of people and yet not dangerous enough to get most world governments to take it seriously. After we lost smallpox and polio I thought any lethal disease for which there was an effective vaccine would be eradicated within a year but COVID-19 looks like it would have turned endemic even without the zoonotic vectors. We have six superbugs more lethal than COVID-19 sitting around in various island fortresses. We had planned to launch them this year but with all the COVID-19 data coming in I'm questioning whether that's really the right way to go. *Primum non boni*. If we release a disease too deadly, governments will stamp it down immediately. We'll kill few people while also training governments in how to stop pandemics. It'd be like vaccinating the planet. We don't want a repeat of SARS.

Director: Good job being quick to change your mind in the face of evidence. What's the status of our AI misalignment program?

Master Roboticist: For several years we've been working on mind control algorithms, but we cancelled that initiative in the face of competition with Facebook. I don't like Facebook. They're not optimally evil. There are many ways they could be worse for the world. But their monopoly is unassailable. The network effects are too great.

Director: Where are our AI misalignment funds going instead?

Master Roboticist: For a while our research was going into autonomous weapons. Autonomous weapons make it easier to start a war since leaders can attack deep within enemy territory without risking the lives of their own soldiers. Predator drones

also cause lots of collateral damage. A drone strike by the Biden administration killed seven children.

Director: If the program's going so well then why stop it?

Master Roboticist: Competition. China is [revamping its military around autonomous warfare](#). We have extraordinary resources, but even we can't compete with the world's biggest economy.

Director: Perhaps we can co-opt their work. Is there no way to start a nuclear WWIII?

Strategos: Starting a nuclear war is easy. It's so easy that our efforts are actually dedicated to postponing nuclear war by reducing accidents. There's more to evil than just causing the greatest harm. We must cause the greatest harm to the greatest number. Our policy is that we shouldn't start a nuclear war while the world population is still increasing.

Master Roboticist: Moreover, a nuclear war would stop us from summoning Roko's basilisk.

Director: How's the AGI project going by the way?

Master Roboticist: Slow. Building an optimally evil AI is harder than building an aligned one because you can't just tell it to copy human morals. Human beings frequently act ethically. Tell a computer to copy a human being and the computer will act ethically too. We need to solve the alignment problem. Alas, the alignment research we produce is often used for good.

Director: Oh no! Can we just keep it secret?

Master Roboticist: Open dialogue is foundational to scientific progress. We experimented with what would happen if we keep our research secret but the do-gooders rapidly outpaced us.

Director: What about fossil fuels? Those are usually a source of villainous news.

Uncivil Engineer: Building solar power plants is cheaper than building coal power plants.

Director: That sounds insurmountable.

Uncivil Engineer: You'd think so. But we got a major news outlet to print a quote about how using coal power plants to mine cryptocurrency for a private equity firm "can have a positive emissions impact if it's run the right way".

Director: You're kidding.

Uncivil Engineer: [I'm not.](#)

Comments on OpenPhil's Interpretability RFP

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

I'm very excited about research that tries to deeply understand how neural networks are thinking, and especially to understand tiny parts of neural networks without too much concern for scalability, as described in [OpenPhil's recent RFP](#) or the [Circuits thread on Distill](#).

I want to give some detail on the kind of research I'm most excited about and why.

(Disclaimer: These are off-the-cuff thoughts. I'm not at all an expert in this area, just someone who cares a lot about alignment and has opinions.)

Why I'm excited about interpretability

I'm very scared of powerful AI doing something deliberately destructive that no one wanted it to do. If we had a reasonably complete story about how a neural network was thinking, then I doubt this failure mode would take us by surprise---we'd be able to see that the model was only behaving well on the training distribution because we were watching, and understand that the model was capable and inclined to cause harm when deployed. I think that even a partial understanding could provide a huge amount of clarity on open disputes about how models will generalize and when they may pose risks.

So the simplest positive case for interpretability is that it gives us a much clearer warning about an impending problem. That may or may not be enough for us to get our act together and fix the problem, but it's at least a start.

Beyond that, interpretability plays a central role in many proposals for training safe AI. For example, we might use interpretability to help identify cases where the model would behave badly, to help ensure that the model never develops deceptive intent, or to help understand what the model knows about the world so that we can better evaluate its outputs. We don't really know which of these approaches will pan out, but the fact that interpretability keeps coming up makes it seem important beyond just a warning shot.

Some of these approaches would put more "strain" on an interpretability method. If we optimize models so that they don't appear to think certain kinds of thoughts, then gradient descent will find a way to hide those thoughts. I don't know if interpretability techniques would ever be strong enough to stand up to this kind of optimization pressure, but I think interpretability is incredibly useful anyway---it still gives you a key warning shot, and many of the potential applications in the training process wouldn't create as much adversarial pressure.

It's possible for models to eventually hide their cognition even if we don't apply gradient descent. But I think this may be extremely hard; that the process of doing so is likely to itself be visible if we are careful; and that even if transformative AI is smart

enough to obfuscate its thoughts we would still learn a huge amount from studying dumber models.

I'm significantly more interested and optimistic about mechanistic interpretability than I was 4 years ago. So from my perspective Chris Olah, who has championed this perspective during the interim, gets a lot of epistemic credit here.

Why I'm not worried about scalability

I think I basically totally agree with the [OpenPhil RFP](#), and in particular think the most exciting work is getting an extremely detailed mechanical understanding of a tiny part of a neural network. If anything I'd lean slightly further in that direction than Chris does in the RFP.

Expanding on that: it seems to me that the current bottleneck for interpretability is that we almost never have a good understanding of what's going on, not that we have good methods that scale poorly. I think people should mostly not worry about scalability while we try to get really good at understanding what's up with small pieces of neural networks.

That's partly an aesthetic judgment about what the conceptual bottlenecks are, but I think there are also some concrete reasons not to worry too much about scalability:

- Worst case, in the long run I think there's a very good chance that it would be feasible to automate the literal steps a human takes in order to understand a neural network, and that doing so will be cheap enough to apply to even the largest models.
- Each time we do this exercise I think we will learn much more about how to do it well and about the general structure of neural networks. I think there's a good chance that if we understand how to do this task well we'll also understand how to write a scalable algorithm.
- *Maybe* if we had many examples of understanding small circuits, and it wasn't getting any easier and we weren't learning anything general, then I'd suggest that we should be focusing on scalability. But instead I feel like this work is at an extremely preliminary stage, where we have very few examples of deeply understanding anything other than low-level vision and could go a lot further even on the examples we do have.
- I have some general intuition that it's very unlikely for this kind of task to be bottlenecked on computational issues---if we're in that world, then in some sense interpretability would be one of the *easiest* parts of ML to work on, and so I would expect extremely rapid progress.

I think "fully understand a neural network" is a good aspirational goal, which I think is also mostly bottlenecked on "deeply understand small pieces effectively."

Comparison to Circuits

The [Circuits thread on Distill](#) is probably the best example of work I find exciting.

I think the most exciting aspect of this thread is the "Artificial Artificial Neural Network" described in [Curve Circuits](#), and in particular the preliminary results on replacing the "natural" curve circuit with their hand-coded version (See "Finally, a

preliminary experiment..."). They consider a set of neurons that appear to be doing curve detection, find that zeroing those activations reduces accuracy by 3.3%, and that replacing them with a hand-written curve detection algorithm recovers half of the loss.

I feel like "understand pieces of cognition well enough that you can replace them without degrading performance" is the right game to be playing, and that the evidence provided by successfully doing the replacement would be *much* stronger than anything else that we think we know about neural networks.

My biggest reservations with that work are:

- The experimental evidence that the artificial circuit works well is preliminary. The sample size is small enough that it could just be noise; I doubt the whole effect is noise, but the real effect size could at least be very different.
- I think that having the artificial circuit work at least as well as the original would be much more informative than being 50% as good; this more ambitious goal also seems tractable with a good enough understanding. It could be the case that zeroing out a neuron is extremely harmful for a network (much worse than retraining from scratch without those neurons), such that even approximating some crude high-level statistics would be 50% as good.
- It seems plausible that "curve detector" is one of the easiest circuits, and that replacing even a single random neuron in an image model would be much harder. I wouldn't want to address that by trying to jump straight to a random neuron, but it's something you'd learn about as you did more examples.

Most of all, I think that they could predictably get much cleaner results by continuing to spend time on the problem---they didn't put much effort into this experiment, and they mention several obvious ways to do much better (e.g. their version of the circuit is in grayscale instead of color).

There were several ways I think that the authors might have been making life unnecessarily hard on themselves:

- They implemented their algorithm using exactly the same architecture as the original, setting individual weights by hand. But I think it would be nearly as good to replace a neuron with an arbitrary (efficient) function of the earlier neurons. This makes it much more plausible that you will recover >100% of the performance and hence have confidence that you understand the high order bits of what the model is doing.
- Chris avoided looking at the model while initially implementing the replacement. This makes a lot of sense as a simple way to reduce the probability of cheating, but in the long run I think you want to be going back and forth, analyzing errors and divergences between your version and the original in order to learn more and more about it. That requires making a more subtle judgment about whether you "understand" the algorithm you've written down, but I think that's kind of in the nature of the game.
- They were looking for motifs and patterns that could allow them to simplify the network and understand bigger pieces of it at once, but I think it would be fine just to get a great grip on e.g. a single orientation of curve detector. (This makes it harder to confirm that you've preserved performance, but it still looks tractable if you do a careful evaluation.) I think the biggest implication is that future work should probably be happy understanding even smaller parts of neural networks since we won't always have so much equivariance.

If I were recommending someone an applied alignment project to get started in the field, a strong candidate would be trying to "finish the job" on curve detectors, before trying to apply the same treatment to another similarly complex neuron. My biggest concern would be that this work would be quite hard and not adequately appreciated, but given increasing interest in alignment and consensus about the centrality of interpretability I'm less concerned about that than I would have been a few years ago.

I'm somewhat more excited about doing these analyses for large transformers trained as LMs. I think there are good reasons to expect the same basic approach to work, and that if anything the existing problem is more likely to be usefully analogous to the long-term problem. That said, I think that there are a few tricky things in generalizing this approach to transformers and for people new to the field it may be better to try to follow more closely with the existing work on vision models (since I expect the core difficulties to be extremely similar).

In general I think that this work is more likely to be tractable for smaller models, and I'm intuitively a bit skeptical of Chris' "[Valley of Confused Abstractions](#)". I'm scared about projects shifting prematurely to larger models because the *most interpretable* parts of the model get easier and easier to understand even while the *least interpretable* parts (where in some sense we should be aiming) are getting harder and harder to understand. That said, I have almost no first-hand experience doing interpretability, and Chris has *much* more experience doing interpretability including with tiny models---so I think there's a good chance that projects on e.g. an MNIST model would just be exercises in clarifying what it really means to e.g. "understand" a linear regression, rather than expecting to get very satisfying results.

Some caveats

I don't really expect to be able to tell "simple" stories about individual neurons in general. I think many of them might be complicated messes involving unfamiliar concepts where a human can only understand one small aspect of a neuron's behavior at a time. And often a neuron will only make sense in the context of a bunch of other neurons.

The RFP talks about polysemanticity as something we'd like to avoid, but I'm a bit skeptical on that point---it seems to me like in powerful systems neurons will often be incredibly polysemantic, and indeed to stretch the concepts such that "polysemantic" doesn't really make sense (consider a transistor in my computer). That said, I also would have intuitively predicted existing models to be quite polysemantic, and so successful monosemantic replacements of neural net circuits suggest that I'm mistaken and maybe things will be simpler than I expected for longer.

Even given highly polysemantic neurons without any simple human-understandable stories, we can still have the goal of writing algorithms we fully understand that achieve the same loss. The feasibility of that project seems like it may be very closely related to the feasibility of "factored cognition:" can we build arbitrarily complex structures in human-understandable ways out of human-understandable parts? For challenging models this is likely to involve significant caveating and clarification about what "understand" means, etc. But to the extent that those issues arise in interpretability, I think it's reasonable to cross that bridge when we come to it and that it's a reasonable domain to explore the same questions that would be important for IDA or other approaches.

(Moreover, from my perspective one of the main obstacles to making progress on factored cognition right now is that we just don't have very compelling example domains where ML systems understand important things in ways we can't. So if interpretability fails for this kind of conceptual reason, then at least we plausibly get a consolation prize of an interesting and plausibly challenging test case for other techniques.)

A Bayesian Aggregation Paradox

In short: There is no objective way of summarizing a Bayesian update over an event with three outcomes A : B : C as an update over two outcomes A : $\neg A$.

Suppose there is an event with possible outcomes A, B, C.

We have prior beliefs about the outcomes $p_1 : p_2 : p_3$.

An expert reports a likelihood factor of $e_1 : e_2 : e_3$.

Our posterior beliefs about A : B : C are then $p_1 \cdot e_1 : p_2 \cdot e_2 : p_3 \cdot e_3$.

$$\begin{array}{ccc} \left(\begin{array}{c} p_1 \\ p_2 \\ p_3 \end{array} \right) & \times & \left(\begin{array}{c} e_1 \\ e_2 \\ e_3 \end{array} \right) \\ \text{Prior} & & \text{Update} \end{array} = \left(\begin{array}{c} p_1 \cdot e_1 \\ p_2 \cdot e_2 \\ p_3 \cdot e_3 \end{array} \right) \quad \text{Posterior}$$

But suppose we only care about whether A happens.

Our prior beliefs about A : $\neg A$ are $p_1 : (p_2 + p_3)$.

Our posterior beliefs are $p_1 \cdot e_1 : (p_2 \cdot e_2 + p_3 \cdot e_3)$.

This implies that the likelihood factor of the expert regarding A : $\neg A$ is $\frac{p_1 \cdot e_1}{p_1 \cdot (p_2 + p_3)} : \frac{(p_2 \cdot e_2 + p_3 \cdot e_3)}{p_2 + p_3}$

$$\begin{array}{ccc} \frac{p_1}{(p_2 + p_3)} & \times & \left(\begin{array}{c} e_1 \\ \frac{p_2 \cdot e_2 + p_3 \cdot e_3}{p_2 + p_3} \end{array} \right) \\ \text{Prior} & & \text{Update} \end{array} = \left(\begin{array}{c} p_1 \cdot e_1 \\ p_2 \cdot e_2 + p_3 \cdot e_3 \end{array} \right) \quad \text{Posterior}$$

This likelihood factor depends on the ratio of prior beliefs $p_2 : p_3$.

Concretely, the lower factor in the update is the weighted mean of the evidence e_2 and e_3 according to the weights p_2 and p_3 .

This has a relatively straightforward interpretation. The update is supposed to be the ratio of the likelihoods under each hypothesis. The upper factor in the update is $P(E|A)$. The lower factor is $P(E|B \cup C) = \frac{P(B)P(E|B)+P(C)P(E|C)}{P(B)+P(C)}$

$$\begin{array}{ccc} P(A|E) & & P(A) & & P(E|A) \\ (P(B \cup C|E)) & \propto & (P(B \cup C)) & \times & (P(E|B \cup C)) \\ \text{Posterior} & & \text{Prior} & & \text{Update} \\ \\ P(E|A) & & \left(\frac{P(E|A)}{P(E|B \cup C)} \right) & & \left(\frac{P(E|A)}{P(E|B \cup C)} \right) \\ (P(E|B \cup C)) & = & \left(\frac{P(E|B \cup C)}{P(B \cup C)} \right) & = & \left(\frac{P(B)P(E|B)+P(C)P(E|C)}{P(B)+P(C)} \right) \\ \text{Update} & & & & \end{array}$$

I found this very surprising - the summary of the expert report depends on my prior beliefs!

I claim that this phenomena is unintuitive, and being unaware of this can lead to errors.

Why this is weird

Bayes' rule describes how to update our prior beliefs using data.

In my mind, one very nice property of Bayes rule was that it cleanly separates the process into a subjective part (eliciting your priors) and an ~objective part (computing the update).

$$\begin{array}{l} \text{Posterior} = \text{Prior} \times \text{Likelihood} \\ \hline \square \square \square \square \quad \square \square \square \square \square \square \square \end{array}$$

Subjective Objective

For example, we may disagree on our prior beliefs on whether eg COVID19 originated in a lab. But we cannot disagree on the direction and magnitude of the update caused by learning that [it originated in one of the few cities in the world with a gain-of-function lab working on coronaviruses](#).

Because of this, [researchers are encouraged to report their update factors together with their all considered beliefs](#). This way, users can use their research for their own conclusions by multiplying their prior with the update. And metastudies can just take the product of the likelihoods of all studies to estimate the combined effect of the evidence.

In the above example, we lose this nice property - **the update factor depends on the prior beliefs of the user**. Researchers would not be able to objectively summarize their likelihood about whether COVID19 originated in a lab accidentally vs zoonotically vs being designed as a bioweapon as a single number for people who only care about whether it originated in a lab versus any other possibility.

Examples in the wild

I ran into this problem twice recently:

1. When analyzing [Mennen's ABC example](#) of a case where averaging the logarithmic odds of experts seems to result in nonsense.
2. In my own research on [interpreting Bayesian Networks](#) as I was trying to come up with a way of decomposing a Bayesian update into a combination of several updates.

In both cases being unaware of the phenomena led me to a conceptual mistake.

Mennen's ABC example

Mennen's example involves three experts debating an event with three possible outcomes, A : B : C.

Expert #1 assigns relative odds of 2 : 1 : 1.

Expert #2 assigns relative odds of 1 : 2 : 1.

Expert #3 assigns relative odds of 1 : 1 : 2.

The logodds-averaging pooled opinion of the experts is $\frac{-\sqrt{2}}{\sqrt{2}} : \frac{-\sqrt{2}}{\sqrt{2}} : \frac{-\sqrt{2}}{\sqrt{2}}$ i.e. equal odds, which correspond to a probability of A equal to $\frac{1}{3} \approx 33.33\%$.

$$\begin{array}{ccccccc} & \overline{\square \square \square \square \quad \square \square \square \square \quad \square \square \square \square \quad \square \square \square \square} & & & & & \left(\frac{-\sqrt{2}}{\sqrt{2}} \right) \\ \square & \left(\begin{array}{c} 2 \\ 1 \\ 1 \end{array} \right) & \left(\begin{array}{c} 1 \\ 2 \\ 1 \end{array} \right) & \left(\begin{array}{c} 1 \\ 1 \\ 2 \end{array} \right) & = & \left| \begin{array}{c} \frac{-\sqrt{2}}{\sqrt{2}} \\ \frac{-\sqrt{2}}{\sqrt{2}} \\ \frac{-\sqrt{2}}{\sqrt{2}} \end{array} \right| \\ \square & \times & \times & & & & \\ \square & \downarrow \text{Expert #1} & \downarrow \text{Expert #2} & \downarrow \text{Expert #3} & & \downarrow \text{Pooled opinion} & \end{array}$$

But suppose we only care about A : $\neg A$.

Expert #1's implicit odds are 2 : 2.

Expert #2's implicit odds are 1 : 3.

Expert #3's implicit odds are 1 : 3.

The pooled odds in this case are $\frac{-\sqrt{2}}{\sqrt{2}} : \frac{-\sqrt{2}}{\sqrt{2}} : \frac{-\sqrt{2}}{\sqrt{2}}$, which correspond to a probability of A equal to $\frac{-\sqrt{2}}{\sqrt{2} + \sqrt{2} \cdot 3} \approx 32.47\%$.

$$\begin{array}{ccccccc}
 & \hline
 & 2 & & 1 & & 1 & \\
 & \square (1+1) & \times (2+1) & \times (1+2) & = (\sqrt[3]{2 \times 3 \times 3}) \\
 & \square \square \square \square \square \square \square \square \\
 & \downarrow \text{Expert #1} & \text{Expert #2} & \text{Expert #3} & \text{Pooled opinion}
 \end{array}$$

We get different results depending on whether we take the implicit odds after or before pooling expert opinion. What is going on?

Mennen claims that this is a strike against logarithmic pooling. The issue according to him is in the step where we take the opinion of the three experts and aggregate it using average logodds.

I think that this is related to the phenomena I described at the beginning of the article. The problem is with the step where we take the relative odds $1 : 2 : 1$ and summarize them as $1 : 3$.

It's no wonder that logodd pooling gives inconsistent results when we aggregate outcomes. Bayesian updating is not well defined in that case!

Interpreting Bayesian Networks

I will not enter into too much detail because [my theory of interpretability of Bayesian Networks](#) is very complex. But it suffices to say that I was getting inconsistent results because of this issue.

In essence, I came up with a way of decomposing a Bayesian update into a series of independent steps, corresponding to different subgraphs of a Bayesian Network.

For example, I would decompose the update over a node with three outcomes A, B, C as the product of the baseline odds of the event and a number of updates.

In my system, I only cared about whether A happened. So I naively summarized each update before aggregating them.

$$O(\text{Event} | \text{Evidence}) \approx (p_1 + p_2 + p_3) \times (e_{1,1} + e_{1,2} + e_{1,3}) \times \dots \times (e_{n,1} + e_{n,2} + e_{n,3})$$

$\square \square \square \square \square \square$	$\square \square \square \square \square \square$	$\square \square \square \square \square \square$
Prior	Argument 1	Argument n

This was giving me very poor results - my resulting updates would be very off compared to traditional inference algorithms like message passing.

It is no wonder this was giving me bad results - it is the wrong way of going about it! Our analysis at the beginning implies that the update should be the average of $e_{1,2}$ and $e_{1,3}$, instead of the sum.

After realizing the paradox, I changed my system to not summarizing the odds of A : $\neg A$ until after aggregating all the updates.

$$O(\text{Event} | \text{Evidence}) \approx \frac{(p_1)}{(p_2 + p_3)} \times \frac{(e_{1,1})}{(e_{1,2} + e_{1,3})} \times \dots \times \frac{(e_{n,1})}{(e_{n,2} + e_{n,3})}$$

$\square \square \square \square \square \square$	$\square \square \square \square \square \square$	$\square \square \square \square \square \square$
Prior	Argument 1	Argument n

Performance improved.

Consequences

I am quite confused about what to think about this.

It clearly has consequences, as illustrated by the examples in the previous section. But I am not sure what to recommend doing in response.

My most immediate takeaway is to be very careful when aggregating outcomes - there is an important chance we will be introducing an error along the way.

Beyond that, the aggregation paradox seems to imply that **we need to work at the correct level of aggregation**. We cannot naively deduce implied binary odds from the distribution of a multiple outcome event.

But what is the right level of aggregation?

When aggregating, the lower factor of the update is a weighted mean of the evidence likelihoods $P(E|B)$ and $P(E|C)$. This suggests that the problem disappears when we impose $P(E|B) = P(E|C)$ for any disaggregation of the joint event $\neg A$ into subevents B and C.

But this condition is too strong. For example, we could base our disaggregation on the observed evidence. For example, if the evidence E can either be Red or Blue we could disaggregate $\neg A$ into the cases where E = Red and the cases where E = Blue. In that case, the condition cannot ever be satisfied, by definition.

We can say that this disaggregation is not a sensible one, and ought to be excluded for the purposes of the condition. But in that case we have passed the buck down to defining what is a sensible disaggregation.

Another approach is to assume that the prior relative likelihood of any aggregated outcomes is uniform, ie $P(B) = P(C)$. In that case, we have that $P(E|B \cup C) = \frac{P(B) \cdot P(E|B) + P(C) \cdot P(E|C)}{P(B) + P(C)} = \frac{P(E|B) + P(E|C)}{2}$.

But then we can no longer chain updates - after applying any likelihood where $P(E|B) \neq P(E|C)$ the resulting posterior will no longer meet this condition.

Pragmatically, it seems like the best we can do if we want to rescue objectivity is to resign ourselves to summarize the updates assuming a uniform prior. That is, by averaging the evidence associated to each aggregated outcome.

This is not enough to correctly approximate Bayesian updating, as we can see in the example below:

$$\begin{array}{ccccccccc} \left(\begin{array}{c} 1 \\ 1 \end{array} \right) & \left(\begin{array}{c} 1 \\ 1 \end{array} \right) & \left(\begin{array}{c} 1 \\ 0.01 \end{array} \right) & \left(\begin{array}{c} 1 \\ 1 \end{array} \right) & & \left(\begin{array}{c} 1 \\ 1+1 \end{array} \right) & \times \left(\begin{array}{c} 1 \\ \frac{0.01}{2}+1 \end{array} \right) & \times \left(\begin{array}{c} 1 \\ \frac{1}{2}+0.01 \end{array} \right) & \approx \left(\begin{array}{c} 1 \\ ? \end{array} \right) \\ | 0.01 | & = | 1 | & \times | 1 | & \times | 1 | & & | 1 | & & | 1 | & \\ \left(\begin{array}{c} 1 \\ 0.01 \end{array} \right) & & \left(\begin{array}{c} 1 \\ 1 \end{array} \right) & & & \left(\begin{array}{c} 1 \\ 1 \end{array} \right) & & \left(\begin{array}{c} 1 \\ 1 \end{array} \right) & \\ \text{Posterior} & \text{Prior} & \text{Refute B} & \text{Refute C} & & \text{Prior} & \text{Refute B} & \text{Refute C} & \text{Pc} \end{array}$$

But I can't see how to do better in the absence of more information.

One key takeaway here is that **beliefs and updates are summarized in different ways**.

$$\begin{array}{ccc} \left(\begin{array}{c} p_1 \\ p_2 \\ p_3 \end{array} \right) & \xrightarrow{p_1} & \left(\begin{array}{c} e_1 \\ e_2 \\ e_3 \end{array} \right) \\ \text{Belief} & \text{Summarized belief} & \text{Update} \\ & & \xrightarrow{e_1 + e_2 + e_3} \text{Summarized update} \end{array}$$

In summary

I have explained one counterintuitive consequence of Bayesian updating on variables with more than two outcomes. This paradox implies that we should be careful when grouping together outcomes of a variable. And I have shown two situations where this unintuitive consequence is relevant.

This is a post meant to explore and start a discussion more than provide definite answers. Some things I'd be keen on discussing include:

1. Is this a documented phenomena? Where can I find more discussion?
2. What does this imply for formulating forecasting questions? Will this result in problems when asking binary questions about events that are multifaceted?
3. What is "the right level" of outcome aggregation for a given problem?
4. Are there other examples where similar issues come up?

I'd be really interested in your thoughts - please leave a comment if you have any!

Acknowledgements

Thanks to rossry, Nuño Sempere, Eric Neyman, Ehud Reiter and ForgedInvariant for discussing this topic with me and helping me clarify some ideas.

Thanks to Alex Mennen for coming up with the example I referenced in the post.

Coordination Skills I Wish I Had For the Pandemic

In the previous post I noted the pandemic as a wakeup call for coordination-capabilities. There are a few different lenses to look at this through.

Today, I want to look through the lens of *skills I wish I had had*.

There are many pandemic-relevant skills that aren't (especially) related to coordination. It'd have been handy to have a strong background in epidemiology, for example. But here I want to focus in skills that I think would be generally useful, and which bear on solving coordination problems in novel circumstances.

Coordination Bottlenecked On Skills

Many day-to-day coordination activities require me to have skills.

For example, I might buy groceries. This is a coordination activity – it's only worth a farmer's time to grow extra food if a middleman is going to buy it and transport it. It's only worth a middleman's time to transport it if a grocery store will stock it. It's only worth the grocery store's time to stock it if people will pay for it. We don't know exactly what things people want. Me buying groceries sends a signal to the economic system to produce more apples and milk.

Participating in this system is a lot easier if I have a collection of skills, such as reading and speaking English, navigating a grocery aisle, doing basic arithmetic so I know how much my apples and milk will cost. It includes soft skills like “the social norms involved with talking to the cashier.”

Those skills didn't come for free. Society invested in me having them. It could have invested in me having different skills (such as martial prowess), which would enable *different* coordination patterns.

During the 2 months of the pandemic, I found myself wishing I had several new skills that I hadn't previously developed. Each skill would have taken me a month or so to really wrap my head around if I were trying hard. The fact that I needed all five at once felt very overwhelming. I was struggling to be functional at all and executing on the skills I already had.

Each skill was something I think would have been beneficial in single-player mode. I have a speculative sense that if multiple people around me had them, it'd have enabled compound returns. If people *reliably* had them, it may have been possible to build more complex systems on top of them.

Knowing what I value

When I go to the grocery store and don't have enough money to buy both apples and milk, I have to decide which one I value more than the other. This usually isn't too

hard – my sense of “do I want apples or milk more?” is driven by short term feedback loops I’m pretty familiar with.

When I decide whether to accept a job at one company vs another one, I often have a much harder time knowing which I value more. Jobs are multidimensional, varying in pay, longterm skill growth, coworker rapport, etc. They are also high stakes, high investment decisions. It usually takes me several days or weeks to figure out which is preferable.

The pandemic threw me into a situation where many core pillars of my life were ripped out at the same time, while friends, roommates and coworkers were all having core pillars of *their* life ripped out at the same time. Values I’d normally think of as sacred, and not to be traded off, suddenly had to be traded off against each other. It also included *how I value the people around me*, and how I related to their values.

I didn’t know how much I valued my life, or what tradeoffs were worth making for it. I didn’t know what sacrifices I was willing to make for the sake of how other people valuing their lives, or their social lives.

Negotiating (and Maintaining Relationships) Under Stress

In the beginning, I tried to think carefully and negotiate with roommates about everything. But within 2 months I was exhausted of that, and my roommates were exhausted of that. And then a lot of what *could* have been fairly simple discussions ended up too painful and annoying to contemplate.

I’m least confident about how to improve at this skill. In Takeaways from one year of lockdown, Mingyuan notes:

> It's way harder to be a good rationalist in stressful situations... Negotiating in emotionally fraught situations is a very difficult skill, and despite all the training they receive in talking about feelings and what-not, being a CFAR instructor does not make you good at this skill (source: almost everyone in my house was a CFAR instructor or mentor).

But one of the central things seem to be “Be aware that you have a negotiation exhaustion budget. Try to have a sense of which things are actually worth negotiating over. Try to refactor complex social situations into simpler ones that require less negotiation.”

Grieving

At the beginning of the pandemic, I didn’t have much experience with grief. By the end of the pandemic, I had gotten a *lot* of practice grieving for things. I now think of grieving as a key life skill, with particular ramifications for coordination.

It might work differently for different people. But for me, grieving is the act of wrapping my brain around the fact that something important to me doesn’t exist anymore, or can’t exist right now, or perhaps never existed.

It contains two steps – an “orientation” step, where my brain traces around the lines of the thing-that-isn’t-there, coming to understand what reality is *actually* shaped like now. And then a “catharsis” step, once I fully understand that the thing is gone. The first step can take hours, weeks or months. For me, the second step tends to go quickly once I’ve fully processed.

You can grieve for people who are gone. You can grieve for things you enjoyed. You can grieve for principles that were important to you but aren’t practical to apply right now. You can grieve for concepts like “all of my friends and roommates can coexist happily.”

Grieving is important in single-player mode – if I’m holding onto something that’s not there anymore, my thoughts and decision-making are distorted. I can’t make good plans if my map of reality is full of leftover wishful markings of things that aren’t there.

I now think of this as relevant for coordination as well – if I’m hanging onto something that’s not real anymore, the distortion in my map also affects people who are trying to negotiate with me and find the least-bad-option available. My clinging becomes their problem.

Grieving is tricky because it’s often unclear when you’re supposed to grieve, and when you’re supposed to fight for something you still care about.

Grieving healthily takes time. But I now think grieving *healthily and quickly* is a skill you can learn. It does, unfortunately, require you to actually experience things that-need-grieving. The biggest things to grieve are (hopefully) rare.

Calibration

There was a whole bunch I didn’t know about the world, which was necessary to make informed choices.

Some uncertainties were about empirical facts in the external world, relating to covid, civil unrest, economic downturn. How likely am I to catch covid, or give it to my friends? How likely are we to die if we do? What exactly *is* civil unrest, and it is a thing I really need to worry about? Will looting increase? Will there be supply chain breakdowns?

Some uncertainties were about *myself*.

Would I be happier if I moved to the countryside for 6 months? Would I reflectively endorse it given my various commitments to friends, coworkers, and significant other?

Is it worth spending more time resolving conflicts between friends or coworkers about how to navigate the pandemic? If we make an agreement, will I turn out to endorse that agreement?

In all these cases, it’d be great to have perfect knowledge. Perfect knowledge is pretty expensive. But I think it’s a more achievable goal to have *calibrated* knowledge – I at least know what I know, and how wide my confidence intervals are.

After a few failed negotiations wherein I couldn't even tell what was worth negotiating for, I decided to [boot up PredictionBook and start making predictions](#), so I could get a sense of my default calibration.

Numerical-Emotional Literacy (Or: “Scope Sensitivity”)

I think the first few skills might be prerequisites for a kind of deep Numerical-Emotional literacy. (I'm not sure, because I do not yet have deep Numerical-Emotional-literacy)

I know how to use a spreadsheet. What I don't really know is how to connect a spreadsheet to my emotions and motivations.

In the first month of the pandemic, my house had been defaulting to “just do total lockdown”, largely because it was conceptually simple. At some point a housemate said “what if we actually used a spreadsheet to make an informed fermi-model of how dangerous covid could be, and reflect on our values, and use that to consider whether we actually need to be this stringent about lockdown?”

And I agreed with that in principle. But... I just couldn't. I was so stressed out. I didn't have a principled way of valuing my life. I didn't trust myself to be able to do a fermi calc that I'd actually believe in. I didn't trust other roommates to do the fermi calc for me. I didn't have space to learn the skill in a way I *would* trust.

But, man, if I *had* had this skill, and if more of my friends had had it, it would have made a lot of things much easier. In particular because it would have meant we could...

Turning Sacred Values Into Trades

I think it often makes sense to have classes of things that you don't trade away, and that you drop everything to fix if they're threatened.

One of the complaints I heard during the pandemic was “I'd be willing to pay some people, like, \$100s or \$1000s of dollars for them to come to in-person meetings, but everyone is stuck in this mode where they're not willing to even consider it.”

I was one of the people stuck in the mode where I couldn't even consider it. This was in large part due to my obligations to other housemates – everyone was burned out from negotiation and thinking about covid-risk.

At the time, I don't think there was much opportunity to improve on the situation. I think it's pretty harmful to pressure people into accepting deals that they don't feel comfortable making. At least in my corner of the social graph, I know that people were earnestly trying their best and operating with zero cognitive slack for months on end. But it left me wishing for a better world, a world where I, and my friends, *already* had the skills of:

- Having a concrete sense of how we valued our life – how much we'd pay for additional life-hours.

- Having a calibrated sense of how dangerous covid might have been (this could include wide error bars while still having a grounded sense of what the distribution would be, given your current information and your track record predicting things)
- Ability and willingness to recognize when things you previously classified as “key cornerstone of your life you don’t trade away” in fact need to get sacrificed, and the ability to do so with minimal trauma.

If I and several friends had started the pandemic with those skills, I think we’d have been in better positions to figure out where we actually disagreed with each other (as opposed to holing up by default). And then, if we actually disagreed about how much we each valued our lives vs social lives vs working-in-person-together, the additional act of “offer each other trades that are win-win” would have been less overwhelming.

Next post: Systems I Wish Were In Place For the Pandemic. AKA Could we have gotten microcovid.org sooner? Could we have had more numerical-emotional-literacy in the groundwater?

A positive case for how we might succeed at prosaic AI alignment

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

This post is my attempt at something like a response to [Eliezer Yudkowsky's recent discussion on AGI interventions](#).

I tend to be relatively pessimistic overall about humanity's chances at avoiding AI existential risk. Contrary to some others that share my pessimism, however—[Eliezer Yudkowsky, in particular](#)—I believe that there is a clear path forward for how we might succeed within the current [prosaic paradigm](#) (that is, the current machine learning paradigm) that looks plausible and has no fundamental obstacles.

In the comments on Eliezer's discussion, this point about whether there exists any coherent story for prosaic AI alignment success came up multiple times. [From Rob Bensinger](#):

I think it's pretty important here to focus on the object-level. Even if you think the goodness of these particular research directions isn't cruxy (because there's a huge list of other things you find promising, and your view is mainly about the list as a whole rather than about any particular items on it), I still think it's super important for us to focus on object-level examples, since this will probably help draw out what the generators for the disagreement are.

In that spirit, I'd like to provide my own object-level story for how prosaic AI alignment might work out well.

Of course, any specific story for how we might succeed is going to be wrong simply because it specifies a bunch of details, and this is such a specific story. The point, however, is that the fact that we can write plausible stories for how we might succeed that don't run into any fundamental obstacles implies that the problem isn't "we don't even know how we could possibly succeed" but rather "we know some ways in which we might succeed, but they all require a bunch of hard stuff that we have to actually execute on," which I think is a pretty different place to be.

One thing I do agree with Eliezer on, however, is that, when you're playing from behind—as I think we are—you play for variance. That means embracing strategies that might not work in expectation, but that have long tails in the positive direction, and I definitely see my picture here as falling into that category.

Furthermore, as one should probably expect with any full roadmap for solving such a complicated problem, there's still a lot left out of my picture—especially my intuitions for why I think each of these steps is actually plausible. I am currently working on a research agenda that will go into a lot more detail on that, but until it's published, it might be best to just think of this post as an overview of what that agenda will look like.^[1]

Alright, without further ado, here's my concrete picture for how we might end up succeeding:

1. We produce an understanding of a simple, natural class of agents such that agents of this form are capable of doing all of the things that we might want a powerful, advanced AI to do—but such that no agents of this form will ever act [deceptively](#).
 - My current best guess for what such a natural class might look like is a [myopic agent](#)—that is, an agent that only cares about its next action rather than the long-term consequences of its actions. I think it is possible to produce a simple, natural description of myopia such that myopic agents are still capable of doing all the powerful things we might want out of an AGI but such that they never have any reason to be deceptive. [\[2\]](#)
 - In the language of [training stories](#), (1) gives us our *training goal*, the mechanistic description of what sort of model we’re trying to produce.
2. We develop some way of determining whether a given non-deceptive model falls into the natural class we developed in step (1). It’s fine for this not to work for all non-deceptive models, as long as the class of non-deceptive models that it works on is large enough to make (3) and (4) go through.
 - Note that we will never rely on (2) working in a situation where we are given an agent that is already deceptive.
 - One way to accomplish (2) might be to develop [worst-case transparency](#) tools that can tell whether the basic structure of a given model is consistent with (1).
3. We develop a training procedure such that, given that the current model being trained falls into the natural class from step (1), additional training will always keep it in that class.
 - If agents from our natural class are capable of being able to deploy the tools we developed in step (2), then one way to accomplish (3) might be to have the training be done by the model being trained given access to tools from (2).
 - For (3) to just work very straightforwardly, it would need to be the case that the set of models that (2) works for is large enough to include any model that can be reached from one step of training starting from a model in the natural class from (1).
4. Using (2), we guide very early training (before the model has the capability to be deceptive) to get some model (with which we can initialize further training from) that falls into the natural class from (1).
 - For (4) to just work very straightforwardly, it would need to be the case that the set of models that (2) works for is large enough to include any model that can be produced early in training before the model has the capability to be deceptive.
 - Alternatively, the natural class from (1) could just be broad enough to include most models at initialization, though I suspect that will cause problems for (3).
5. Combining (3) and (4), we get an inductive guarantee that we can produce models that fall into our natural class. Because in (1) we constructed our natural class to be sufficient for any tasks that we might want our AI to do, we can now train non-deceptive AIs on any task that we might want them for.

- For the training process in (5) to be competitive, we also need (3) and (4) to not be so resource-intensive that they are substantially harder than training an unaligned model.
 - In the language of [training stories](#), the inductive argument here in (5) is our central *training rationale* for why we'll get a model that satisfies the training goal from (1).
6. Given a powerful and non-deceptive AI produced from (5), we use standard red-teaming (e.g. testing on lots of examples) to find places where the model fails and retrain using (3) until the model looks like it's doing the right thing.
- Because we know that our model is non-deceptive from (5)—and since all of our retraining is done via (3)—the fact that the model looks like it's doing the right thing should give us a real guarantee that it'll actually do the right thing in similar situations, since we know it won't just be pretending to do the right thing.
7. We ensure that the leading AI lab uses (5) + (6) to produce their most powerful and advanced AI systems. By being first, they are able to set the standard for how training powerful machine learning models should generally be done.^[3] Because of [the strong tendencies for AI labs to copy each other's successes](#), other labs also use (5) + (6) to train their powerful and advanced AI systems, ensuring that all of the most powerful AIs in the world are aligned.
- Though these other labs might scale (5) + (6) further, as long as (3) is robust to scale, such systems should stay aligned.
 - Though I think that the forces pushing for homogeneity of AI training processes across labs are strong, once the set of labs with the capability to build misaligned AI systems gets large enough—e.g. once it includes all the small labs too—one of them is bound to break that homogeneity. Thus, there is a period of vulnerability after (7) and before (8) where smaller labs might not follow (5) + (6) and instead build misaligned AI systems.
 - Even if that does happen, however, since it's only small labs with limited capabilities building misaligned AI in a world that already contains aligned AIs built by much larger and more capable labs, it should be quite difficult for such misaligned AI systems to actually destabilize such a world.
8. We use the AI systems from (7) to help us design the next round of powerful and advanced AI systems and develop techniques to end the period of vulnerability.
- I won't say too much about exactly what we would do here, mostly because it's not a problem that we have to solve before we actually get the powerful aligned AI systems to help us solve it, so it's mostly not a problem that I think we need to focus on right now.

If I had to guess what the hardest part of the above picture will be, I'd probably guess (2),^[4] which is why I'm so excited about [Automating Auditing](#) as a way to start making progress on (2) now. That being said, I don't think there are any fundamental obstacles to solving (2)—(2) very explicitly doesn't require us to be robust to deceptive models or even be able to tell whether (1) holds for all non-deceptive

models, both of which I think would run into fundamental obstacles, but which we don't have to do.

1. If you want access to an early draft of my agenda, message me privately and I might send it to you, though it's still likely to change a lot before it's released. [←](#)
2. I think it is possible for a myopic agent to still be capable of solving problems that involve non-myopic reasoning (e.g. be a good AI CEO). For example, a myopic agent could myopically simulate a strongly-believed-to-be-safe non-myopic process such as [HCH](#), allowing [imitative amplification](#) to be done without ever breaking a myopia guarantee—alternatively, [AI safety via market making](#) lets you do [AI safety via debate](#) without breaking a myopia guarantee. In general, I think it's just not very hard to leverage careful recursion to turn non-myopic objectives into myopic objectives such that it's possible for a myopic agent to do well on them—without breaking the guarantees that ensure that your myopic agent won't be deceptive (as a concrete example of what a myopic agent that is capable of doing well on such tasks without ever having any reason to be deceptive might look like, consider [LCDT](#)). [←](#)
3. For an example of what “setting the standard for how training powerful ML models should be done” might look like, consider how once the basic training paradigm of “train massive self-supervised transformer-based language models” was introduced, it was aggressively copied across the field and became the standard for all language-based AI systems. [←](#)
4. Second place for hardest step would probably be (7). Definitely a hard step, but I think that the claim that we mostly only have to persuade the frontrunner makes this not a fundamental obstacle—persuading one organization of one thing is an achievable goal, persuading every person doing AI everywhere would be a fundamental obstacle. [←](#)

Covid 11/25: Another Thanksgiving

For [my first Thanksgiving post](#), which somehow was both a year ago and almost a year into the pandemic, I gave thanks at the end of the post. With all that's happened, it seems more appropriate to do this at the *beginning* of the post instead this year. So I'm going to do that, and deal with this week's news – the rise in cases, the lockdowns in Europe and all that other stuff – later.

I'll be working from the previous version, and including *almost* everything from last year, plus some additions. It is crazy *how little* this needed to be modified after an entire year, and *how many of these are still in the present tense* when they could so easily have been in the past tense. Time goes by, so slowly. Keep pushing that rock up that hill, everyone.

So here we go, 2021 edition.

To all the health care workers. Thank you.

[To all those maintaining the supply lines](#). Thank you.

[To Ryan Peterson](#), who worked to get our containers moving again, even if it didn't end up having as big an impact as we might have hoped. Let this [inspire future efforts](#) across the land. Thank you.

[To those who worked to manufacture the vaccines as quickly as possible](#). Thank you.

To Pfizer, Moderna, Johnson & Johnson and AstraZeneca/Oxford. Thank you.

To all those working on all the other vaccine candidates, even if they didn't work out. Including Sputnik in Russia, and the one in China. And anyone working on all the other treatments, or running any of the experiments or studies, again whether or not your particular effort paid off. Thank you.

To all those working on or expanding capacity for or administering or fighting for the right to do Covid-19 testing, especially rapid testing. Thank you.

To all the essential workers, no matter whether we still call you that. Thank you.

To you, the reader, for being here. Thank you.

To all the commenters, yes all of them. Thank you.

To those who subscribed (Substack version of this post is [here](#)). Thank you.

To all those who have thanked me for doing these columns. You keep me going. Thank you.

To everyone who helped push masks, better masks, Vitamin D, zinc, fluvoxamine, airborne transmission, doing things outdoors, proper ventilation and other key information when official sources were saying otherwise. Thank you.

To Robin Hanson in particular, without whom I would not have felt free to start writing these columns. Thank you.

I'd also single out Tyler Cowen and Alex Tabarrok for their work at Marginal Revolution, which has provided lots of useful thoughts and information. And for the rapid grants. Thank you.

To the Covid tracking project, the Covid machine learning project, and all other similar efforts. You're gone now, but you provided vital information when we needed it most. Thank you.

To the financial security and opportunity to keep me and my family safe, and give me the necessary time to write this column each week. And to the freedom of speech to say what I think each week. Thank you.

To China, South Korea, Japan, New Zealand, Australia and all the other places that, at for a time, beat the pandemic. You showed us it can be done. Regardless of what happened later, you all decided to do a hard thing that mattered, and made it happen. Thank you.

To everyone working to reform or contain the damage from the FDA, CDC, WHO or any other member of the Delenda Est club, or otherwise find ways around regulatory barriers. Thank you.

To the owner of the (still villainous) New England Patriots, who flew a jet in to get protective equipment to health care workers. And everyone else who did what was necessary in the face of banditry, piracy and regulatory obstruction to get people what they need. Thank you.

To those who stood up and called such barriers and enemy actions by their right names. Thank you.

To every politician and public figure and corporate leader, or anyone else, who called upon people to take precautions and also accepted skin in the game and practiced what they preached. Thank you.

To those politicians who did the best they could to help people, based on their model of what would physically help, even if I disagree with that model. Especially strong thanks to those who distributed vaccines and booster shots as widely and quickly as possible in defiance of the FDA and CDC's recommendations. Thank you.

To those who, instead, had a physical model and then called upon us *not* to take unneeded precautions, and also practiced what they preached. Thank you.

To everyone doing what they need to do to keep themselves, their families and friends and their communities safe. Thank you.

To everyone doing what they need to do to keep themselves, their families and friends and their communities sane and thriving through all of this. Thank you.

To my in-laws, who helped us get out of New York City in March and have been invaluable keeping things going out here in Warwick. And to our kid's nanny, without whom disaster would have rapidly ensued. And of course to my amazing wife Laura, for far too many reasons to list here. You all went above and beyond again this year. Thank you.

To my father Solomon, who did his part in all this to help make things better however he could, even if I can't talk about it on the internet. And for teaching us a legit immunology class over zoom, and helping me understand the science whenever I needed it. And for keeping sane through everything that has happened in that tiny apartment. And for continuing to help me understand things that would otherwise have gone over my head or taken me far too many hours. Thank you.

To my former temporary home in [Warwick, New York](#), thank you. You have exceeded almost all of my expectations. This place is highly underrated.

To my cofounder Kathleen Breitman, and my coworkers Alan Comer, Brian David-Marshall, Corey Burkhardt and Paulo Vitor Damo da Rosa, along with many others, as we continue to fight to make the game Emergents a reality. The game is real and the Alpha is beginning. It's been a long road but we can see the finish line now. Thank you.

I'd also like to once again thank India, clarity, disillusionment, consequence and silence. Especially clarity. But screw frailty.

Let's not forget the internet. In particular, Amazon, Instacart, Google and Netflix. Couldn't have done it without you. Thank you.

And also to basically everyone anywhere who modeled the world and is now doing a thing to try and make the physical world better, regardless of whether I think their particular approach is misguided or nonsense. To all the schmucks who think for themselves. To all the live players. You rock. Thank you.

Finally, to everyone I'm forgetting. Thank you too. Comments to thank those I missed are encouraged.

Executive Summary

1. Europe is entering lockdown.
2. Cases continue to rise.
3. [Paxlovid remains illegal.](#)

The Paxlovid news got its own update yesterday.

Let's run the numbers.

The Numbers

Predictions

Prediction from last week: 654k cases (+18%) and 7,500 deaths (+1%).

Results: 593k cases (+7%) and 7,566 deaths (+2%).

Prediction for next week: 620k cases (+3%) and 7,240 deaths (-5%).

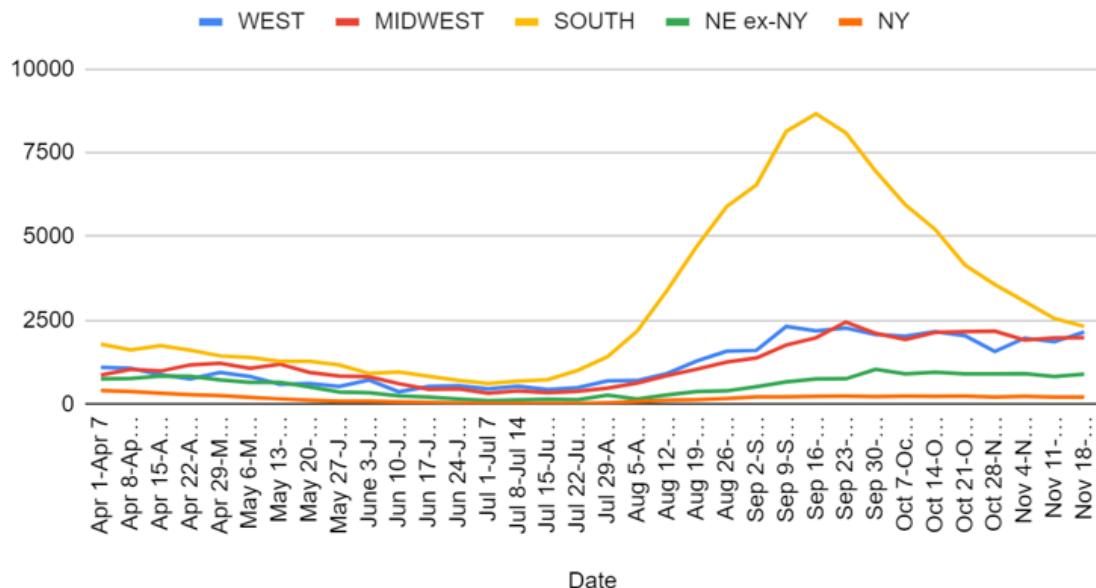
This prediction is imposing a -10% penalty for Thanksgiving, as we only get to count *reported* deaths and cases rather than the actual situation. A lot of the uncertainty here is about how much the holiday will deter testing and reporting, including reporting of deaths.

I would presume that Thanksgiving itself has only a small effect on the course of the pandemic, and also presume that this effect won't yet show itself in next week's data, or will be overwhelmed by the reporting drop.

Deaths

Date	WEST	MIDWEST	SOUTH	NORTHEAST	TOTAL
Sep 30-Oct 6	2076	2113	6955	1261	12405
Oct 7-Oct 13	2031	1929	5949	1142	11051
Oct 14-Oct 20	2166	2143	5213	1183	10705
Oct 21-Oct 27	2044	2163	4151	1148	9506
Oct 28-Nov 3	1575	2177	3572	1115	8439
Nov 4-Nov 10	1970	1914	3068	1140	8092
Nov 11-Nov 17	1860	1978	2558	1031	7427
Nov 18-Nov 24	2155	1985	2319	1107	7566

Deaths by Region



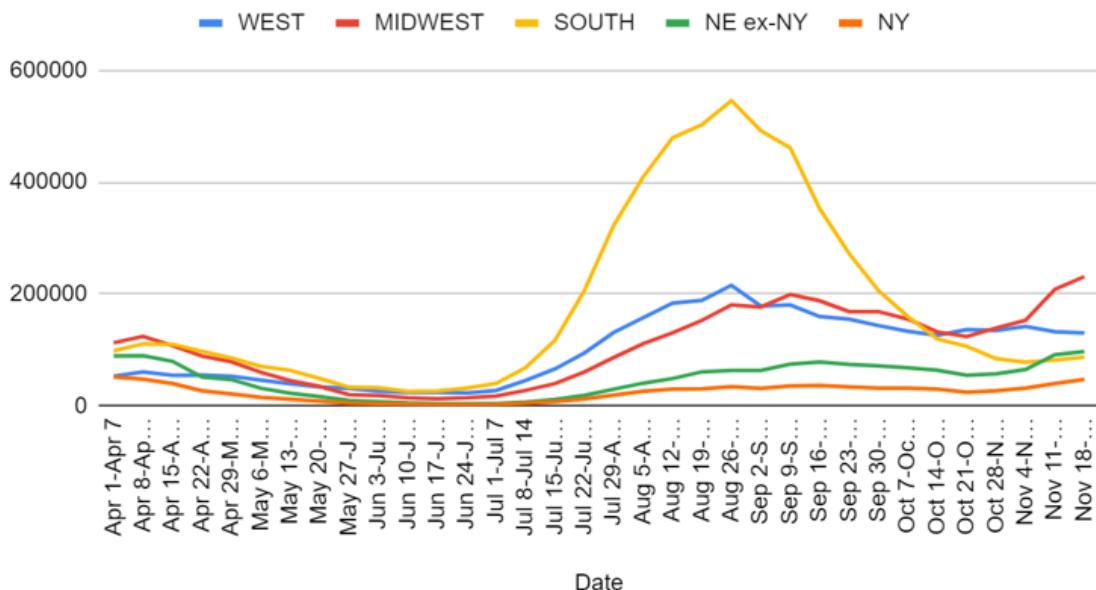
A total of 7,566 people died this past week. Since Pfizer's announcement of results on November 5, about 20,000 people have died in America of Covid-19.

With cases rising, deaths will follow until the approval and widespread use of Paxlovid. At which point they will continue to follow at a much slower rate.

Cases

Date	WEST	MIDWEST	SOUTH	NORTHEAST	TOTAL
Oct 7-Oct 13	133,279	155,015	159,573	99,553	547,420
Oct 14-Oct 20	126,196	132,200	119,798	93,784	471,978
Oct 21-Oct 27	136,528	123,913	106,223	79,364	446,028
Oct 28-Nov 3	134,871	139,171	84,378	84,200	442,620
Nov 4-Nov 10	142,165	152,841	78,090	96,666	469,762
Nov 11-Nov 17	132,626	208,488	81,762	131,285	554,161
Nov 18-Nov 24	130,118	231,105	87,119	144,452	592,794

Positive Tests by Region



Cases are up, including unexpectedly up in the South, but these numbers are good news as reasonable expectations after last week were for things to get worse faster than this. Given deaths were right on path, it's unlikely the holiday had an impact on reporting yet, so this is a large reduction in my estimate of how rapidly things are getting worse.

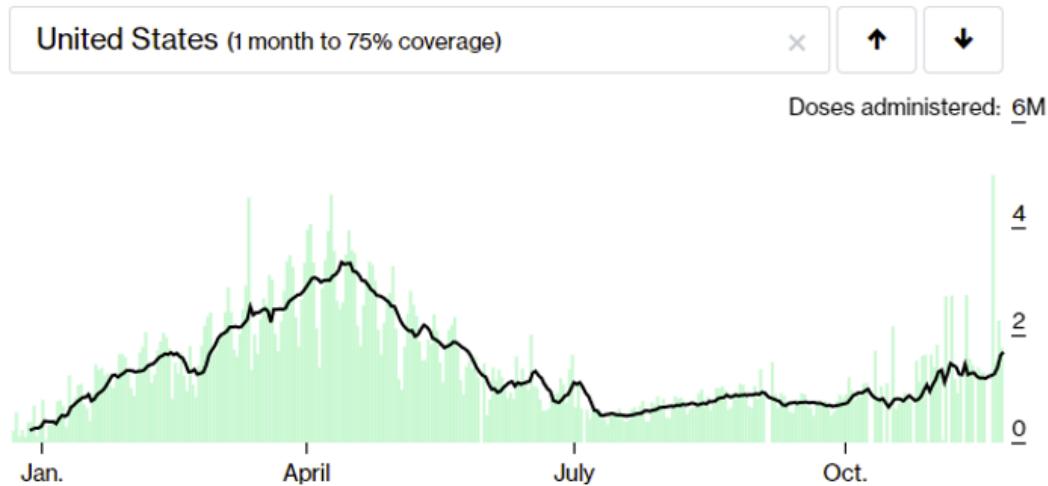
This is especially good news because the situation in Europe has escalated so quickly, and because they have responded with lockdowns. That raises the chance that politicians here might think further restrictions, up to and including lockdowns, might be a good idea when they see the wave continuing, and we very much need to avoid that for many distinct reasons.

Vaccinations

Bloomberg still calls this the 'path to immunity' but at this point it's clear that things do not work that way on the national level. They only work that way for an individual. It's still excellent news to see the rise in vaccinations, even though it is driven by less important child vaccinations rather than adults changing their minds.

The Path to Immunity in the U.S.

In the U.S., the latest vaccination rate is about **1,692,980 doses** per day, which includes **571,464 people** people getting their first shot. At this pace, it will take another **1 month** until **75%** of the population has received at least one dose.

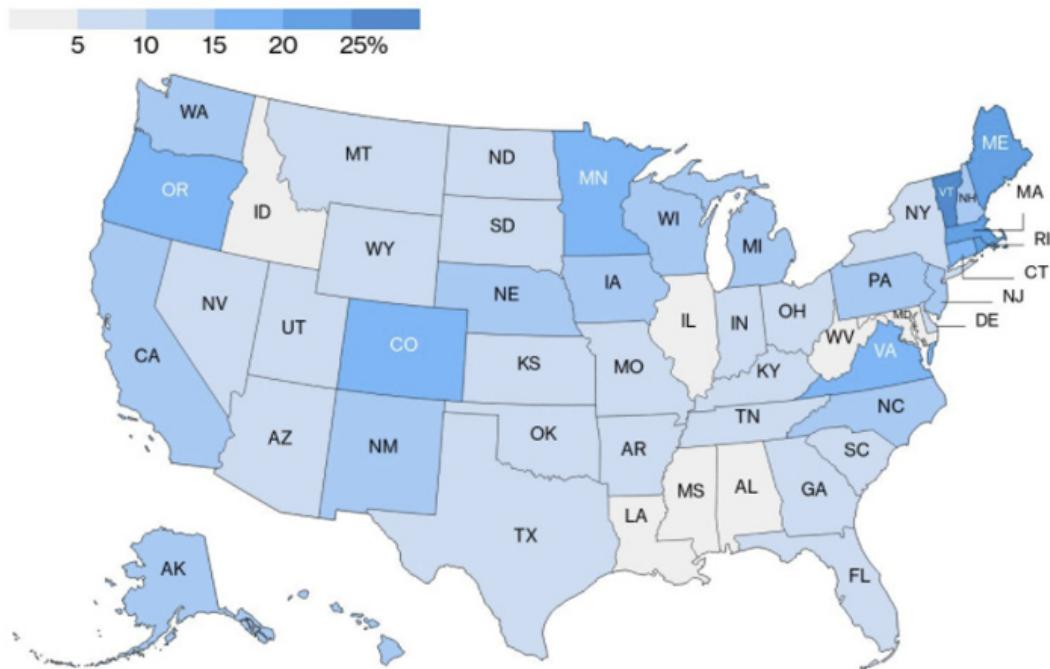


Jurisdiction	Doses administered ▼	% of population					Daily rate of doses administered	Supply used
		per 100 people	given 1+ dose	fully vaccinated	booster dose			
U.S. Totals	454,447,737	136.9	69.7	59.1	11.3	1,692,980	79.5%	
California	58,715,000	148.6	78.2	62.8	11.7	225,747	83.4%	
Texas	36,862,434	127.1	63.7	54.5	9.2	137,055	75.2%	
Florida	30,501,707	142.0	71.3	61.0	11.2	94,186	79.1%	
New York +	29,477,691	151.5	77.5	68.2	9.0	113,449	82.9%	
Pennsylvania	18,031,448	140.8	80.8	58.0	7.4	5,058	76.2%	
Illinois	17,604,565	138.9	68.7	61.4	13.5	65,974	81.9%	

A child vaccination rate graphic showed up in my inbox so here it is:

Vaccine Rates in Young Children

States are progressing at different rates vaccinating kids ages 5 to 11



Source: U.S. Centers for Disease Control and Prevention

Note: Idaho, Maryland have incomplete data for 5-to-11 group. CDC data for other states may lag what states report themselves.

Bloomberg

In total, the CDC showed that 2.84 million young children, or 10% of that population, have received at least one dose of a Covid-19 vaccine.

Booster Boosters

The story of Covid-19 boosters, which have now been given to ~11% of the adult population, seems to broadly be something like this:

1. Boosters seem like an obviously good idea.
2. Biden declares we're going to offer boosters.
3. FDA/CDC get upset, demand that we [respect their authority](#).
4. FDA/CDC begrudgingly approve boosters, but only for the sufficiently vulnerable/worthy of being protected from Covid-19.
5. That [technically includes most people](#), but most people don't know how the rules work.
6. A lot of people realize that the technical rules don't matter at all, what you have to do to be worthy is check a box.
7. [States start deciding this is stupid and opening up boosters to all](#).
8. It turns out that yes, they can do that.
9. [FDA/CDC fold and agree to boosters for all](#).
10. They draw a distinction between who 'should' get a booster, and who graciously 'can' get a booster.



Nate Silver ✅ @NateSilver538 · Nov 17

...

It was pretty obvious all along that we'd eventually approve Booster Shots For All and I'm not sure what we accomplished by hedging except for probably making the winter wave mildly worse.

[Here's a FiveThirtyEight discussion of the situation](#). Which includes this:

betsy: I agree with you, the communication challenge does not help. ... I keep thinking about that [Kaiser Family Foundation poll](#) saying 4 in 10 vaccinated adults don't know if they're eligible for a booster.

maggie: Hell, I've been confused about that.

cwick: It sounds like all three of us are in that 40 percent! And maybe MORE people shouldn't know if they need it.

It also includes this, which seems like a good summary of *one framing* a central thing happening:

Listening to the advisory-committee meetings, I noticed that there seems to be this tension between the scientific experts who want to make robust evidence-based decisions — and the sense that, here in the U.S., [our overall pandemic strategy](#) is basically “vaccinate our way out of the pandemic.”

When I hear the term ‘robust evidence-based decisions’ in contexts like this I mostly interpret it as code for something one might *slightly* uncharitably call ‘this would be a Potentially Blameworthy Action and counts as Medical, and thus requires an isolated demand for rigor, and must prove itself to be overwhelmingly correct based only on properly formatted evidence gathered in Officially Approved ways.’

[This thread covers the ACIP meeting about the topic in detail.](#)

Did you know boosters work? As in hot damn [check out this chart](#)?

Vaccine efficacy during blinded follow-up period by Subgroup
Evaluable Efficacy Population
Without Evidence of Infection Prior to 7 Days after Booster Vaccination

	BNT162b2 (30 µg) N=4,695 n	Placebo N=4,671 n	RVE (%)	(95% CI)
Overall	6	123	95.3	(89.5, 98.3))
Age	16-55 years of age	3	96.5	(89.3, 99.3)
	>55 years of age	3	93.1	(78.4, 98.6)
Sex	Male	70	94.3	(84.8, 98.5)
	Female	53	96.5	(86.7, 99.6)
Race	White	100	95.2	(88.4, 98.5)
	Black or African American	13	100.0	(68.0, 100.0)
	American Indian or Alaska	4	100.0	(-59.2, 100.0)
	Asian	3	69.4	(-280.7, 99.4)
	Multiracial	2	100.0	(-395.3, 100.0)
Ethnicity	Hispanic/Latino	19	94.8	(67.5, 99.9)
	Non-Hispanic/Non-Latino	104	95.4	(88.9, 98.5)

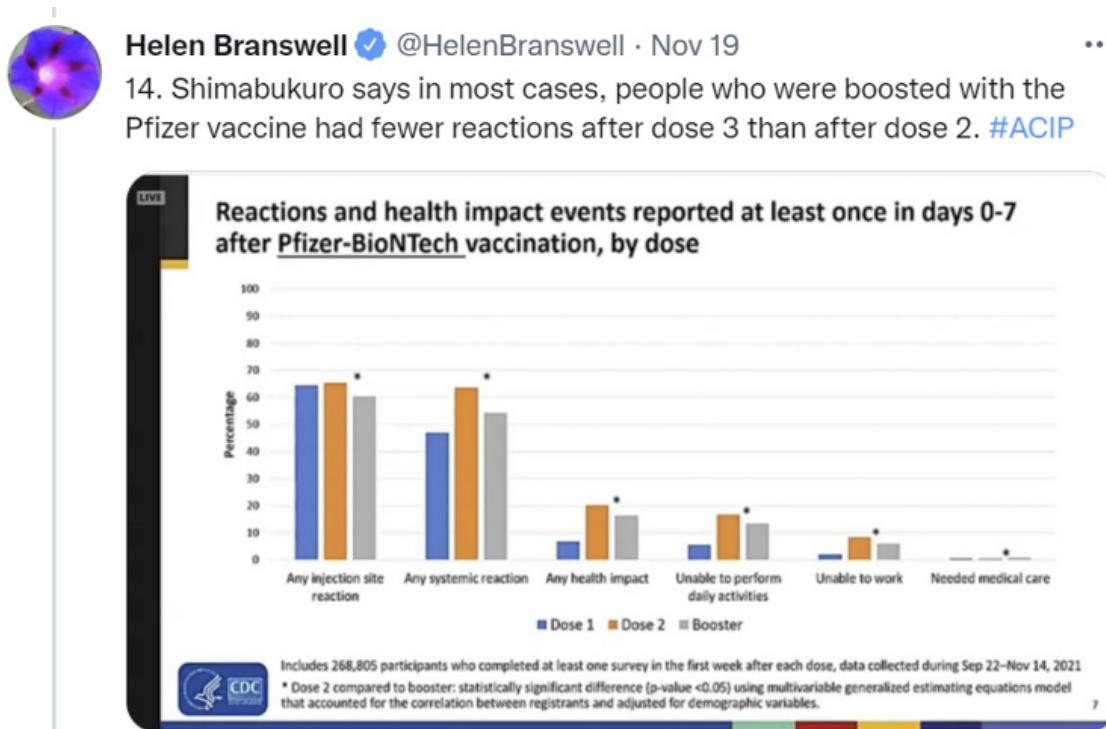


Worldwide Research, Development and Medical

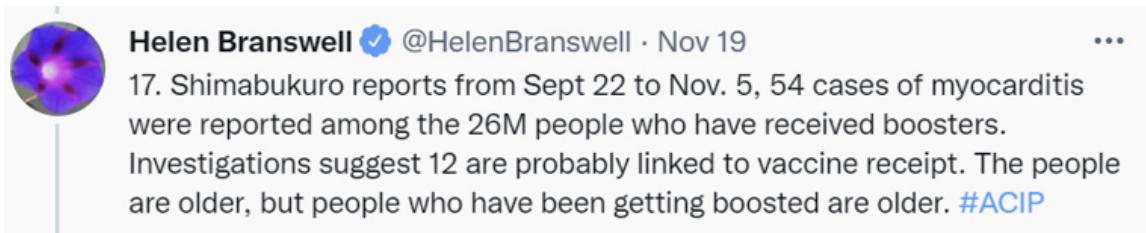
19

I mean, yes, you did know, but a reminder, and especially that the comparison is to an *already vaccinated* control group, seemed worthwhile.

For those wondering about the magnitude of short term side effects, here's the data for Pfizer:



Side effects that make people paranoid watch, why are we still discussing this endlessly and potentially crippling our civilization edition:



It seems that one core argument was that they would have *liked* to reserve prevention of Covid-19 for only those sufficiently worthy, but if you do that the sufficiently worthy aren't sure if they're worthy enough and miss out, so we'll have to compromise and let everyone prevent Covid-19?



The new motions allowing everyone to get a booster passed 11-0, and also were noted to ‘have the support of all states’ so that’s kind of another 50-0 vote. Either this is a [The Death of Stalin](#) situation where once it’s going to pass everyone must vote for it, which wasn’t true for past votes, or we managed somehow to hold out on this until we had universal agreement, in which case it’s an interesting question what took this long.

A lot of the problem seems to be arguments over ‘may’ versus ‘should.’ Right now that seems like a distinction without a difference, but [some are annoyed that we're not telling young people the obviously correct thing that they should get a booster](#), while [others are concerned that if the word should is used it could lead to mandatory boosters](#). Once again, that which is not mandatory is forbidden and that which is not forbidden is mandatory and all that. The flip side is ‘if we don’t use the word should then the young people won’t get the vaccine.’

Plus there’s an unsaid contradiction of the argument that *of course* we should be making vaccine decisions in part based on their social benefits, and we’re going to blame insufficiently vaccinated folks for the pandemic and threaten insufficiently vaccinated populations with both individual and collective punishments, with the fact that officially all decisions can only be made on the basis of individual risks and benefits under extreme risk aversion.

Vaccine Mandates

If you want someone to do something you can:

1. Reward doing it.
2. Punish not doing it.
3. Force them to do it.

Sufficiently large rewards and punishments constitute force. If you are going to use such force, it is better to be explicit that you are doing this, for several reasons.

1. It’s important that we live in a world where we know what’s going on, let the facts enter common knowledge and people call things by their right names. [Here's Taylor Swift taking ten minutes to say that](#) in a different context.
2. [Choices are bad](#). When you give people the illusion of false choice, you’re pretending you’re not forcing them to do it. That makes them feel bad about choosing to do it and agonize about the ‘decisions’ being made and what has been lost, and makes them feel like it is in a way ‘their fault’ or their choice.
3. There’s a lot of lost time and stress and fighting over exactly what’s happening, which would be avoided if you came out and said what you were doing.
4. People can see what you’re doing, and that you’re lying about it. Which pisses them off, as many of my commenters have pointed out in various ways.
5. People can also see that you consider your authority illegitimate. If you *could* mandate the thing outright, you would have, but you didn’t, so you can’t. Which means you lack the legitimate authority to do the things you’re doing. If the authorities tell us they are illegitimate, why shouldn’t we believe them?

Thus, I am far *more* sympathetic to an actual full-on outright mandate than I am to doing various end runs like using OHSA. That particular end run ran into legal problems, and [any enforcement seems to have been abandoned](#) at least for the time being. Enforcement was never the point, since the authority was illegitimate. The point was to give employers cover to impose mandates they wanted to impose anyway.

I am also far more sympathetic to requiring vaccinations than I am to requiring lots of other restrictions that I consider much bigger costs, much bigger violations of liberty (as in, *you can't go outside*) and which are much less effective.

Do, or do not.

Last week regions of Austria instituted a lockdown *on the unvaccinated*. Either get vaccinated, or you can't leave your house without an Officially Approved Excuse. That seems like it is clearly over the line. Then [two regions extended it to the unvaccinated](#) as well, then [the country went into full lockdown](#).

Alongside the lockdown, [Austria has come out and said it](#). Vaccines for everyone. To me, this obviously comes *well before* issuing a nationwide lockdown and telling people they can't go outside except for daily needs, with no end in sight. To me, if you issued a lockdown, it means you *should have already* issued a full vaccine mandate earlier.

I would not be issuing a lockdown regardless of how bad things get, and I'd be leave-the-country level furious if I lived there, but it would be that much crazier without a full vaccine mandate. I do understand their position, especially given Paxlovid. More on the situation there in the next section.

[The Army bans the unvaccinated from reenlisting, being promoted or other 'favorable personnel actions.'](#) That last one risks giving the game away. The full deadline for soldiers to get vaccinated is December 15.

If you get vaccinated, The New York Department of Corrections will generously let you have facial hair. [Reactions differed](#).



Amorette Miller (she/her)
@Amorettemiller



Now that's an incentive!



Assemblyman Joe Angelino
@JosephAngelino

...

Sounds like a negotiable working condition. Also, what's the incentive for people who can't grow facial hair? ie women officer.

The incentive is you get to be protected from Covid-19? What about guards who didn't *want* facial hair? What about the ones who could use it as an excuse before and now can't, and are worse off because choices are bad? Not everything has to work on everyone in order to be a good idea? Have we tried paying cash? Have we tried paying *more* cash? Have we tried you're among people who are there at gunpoint so it's kind of a job requirement, thanks? From context I'm guessing no.

The Return of the Lockdowns

As of Sunday morning, as per Bloomberg, here's where Europe was:

Pandemic Fallout

New cases are pushing countries to introduce more stringent measures

■ Lockdown ■ Partial lockdown ■ Restrictions for unvaccinated ■ Mandatory teleworking
■ Covid pass/certificate ■ Mask mandate ■ Other restrictions



Source: Bloomberg

Note: Covid pass/certificate typically means vaccination/recovery/test required to enter certain venues, exact measures vary by country

Bloomberg

It has since spread considerably further. The examples below are doubtless incomplete.

Austria is in full lockdown, and I understand their position. At the time the lockdown was instituted, Austria's daily case count was unprecedented, over one person in a thousand each day, and rising rapidly.

[Slovakia has locked down. Italy is considering locking down the unvaccinated.](#)

[Some signs of resistance, warning shots fired at protesters](#) in Rotterdam, at least two people shot.

[Germany, where things are not as bad but where cases are also higher than ever before and rising rapidly, is calling this a 'national emergency'](#) and [has now also pulled the trigger on a full lockdown.](#)

What Germany *doesn't* seem to be doing is treating this as a sufficiently urgent emergency to ensure that the unvaccinated have reasonable ways to get vaccinated, as per [this comment from last week's post](#).

German here. I live in this weird reality where all the media are scapegoating the unvaccinated, yet when I walked into a doctor's office last week fully expecting to get a

shot in my arm right away, they told me what they had already told me in summer: "No, you have to get on a waiting list of about 2-3 months." Quickly looked up online if there's anything else officially available in my city, also no. It seems to me that if you want to scapegoat the unvaccinated for all your pandemic mishandling and overreaction problems, you should at least, ya know, offer them vaccines.

This isn't a supply issue. Others replied that appointments are easily available in Berlin, but that's several hours away. Ensuring quick and easy vaccination appointments does seem to me like the least you could do if you're going to be declaring national emergencies and scapegoating the unvaccinated.

So more detail from another comment, mumble mumble something incentives something something not paying enough money mumble something revealed preferences:

Only German sources (see below) and anecdotal experience:

1. My roommate was in the same position as Anon but got an appointment for his first dose in less than two weeks (was there last week).
2. Regarding incentives: Doctors got 20€ per administered shot, health minister raised it last week to 28€. I don't know if it's financially viable to do vaccinations for doctors. The doctors I know (family friends) are giving out vaccinations, but the majority of their work time is for "normal doctoring". Waiting times around 1-2 months aren't uncommon at regular doctors.
3. In my city, Aachen a medium sized city, there is one place where everyone can get vaccinated (1./2./3. shot) without appointment. It's open 12:00-20:00 thought the waiting times can get quite long after regular business times. (Its in the shopping center where i buy my groceries, so i see the queues several times a week. There are 170 of these centres in Germany.
4. Also there are "vaccination busses", mobile opportunities to get vaccinated without appointment in Aachen and surrounding cities. Around 700 of these are driving around Germany.

That does not sound to me like the appropriate response to a 'national emergency.'

Yet they claim to be taking the matter seriously, [as you see here](#).



Daniel Eth

@daniel_eth

...

I'm all for vaccine mandates, but the penalty here seems a tad too extreme

THE LOCAL 



Germany's news in English

Germans will be 'vaccinated, cured or dead' after winter, minister claims

If you look at deaths rather than cases, even if you take into account the time lags, things look much less dire. Vaccinations work, and also our procedures have improved.

There's still quite the large surge, and [vaccination does not seem to be a sufficient defense](#):



Dr. Eli David @DrEliDavid

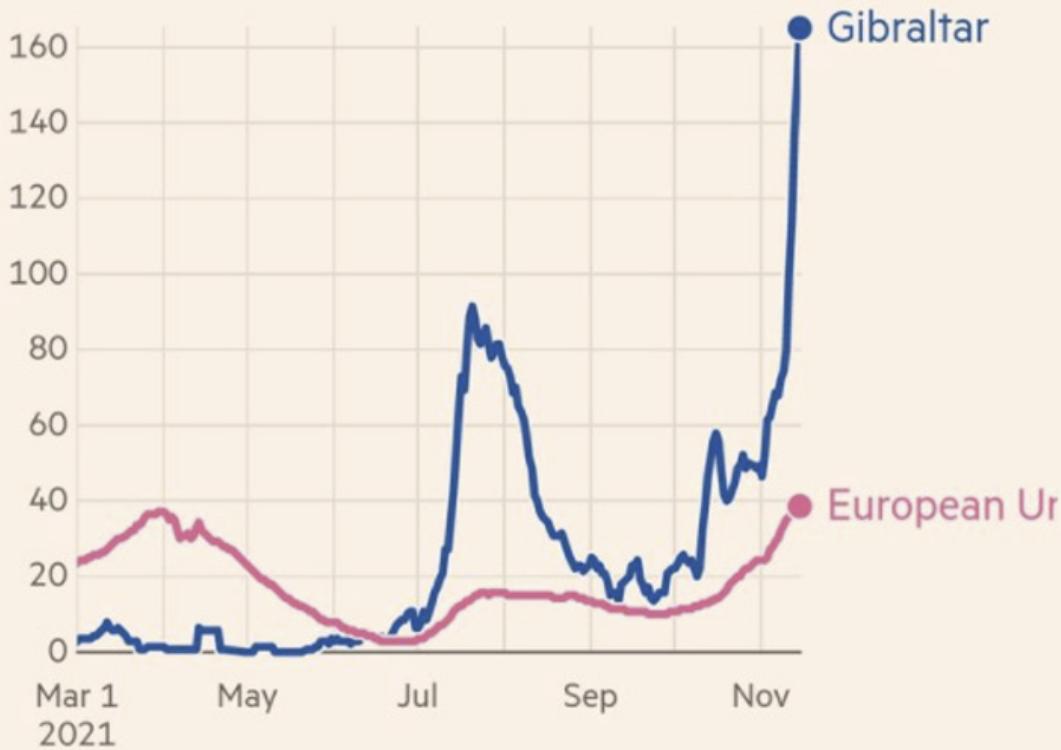
...

In Gibraltar 🇬🇮, the most vaccinated place on Earth, 100% of the population is vaccinated and 40% with booster, but:



New confirmed cases of Covid-19 in Gibraltar and European Union

Seven-day rolling average of new cases (per 100k)



There's a lot of deeply stupid takes about how this means vaccines 'don't work' but it does mean that they by themselves, including with a 40% booster shot rate, are insufficient at containing infections. Which is a much less scary prospect than without the vaccinations, but still matters, even if everyone is vaccinated and therefore there won't be that many deaths.

The replies trying to ‘support’ vaccinations are pure clown-makeup-meme, you see vaccinations were never supposed to *prevent infections* or anything, ‘the basics of science’ means repeating the Party Line, can such people *please* stop helping.

 **Brigid620** @Brigid620 · Nov 19 · ...
Replies to [@DrEliDavid](#)
What's your point? The vaccines were never meant to stop all COVID cases they were meant to keep hospitalizations, ICU capacity and deaths down. We will all eventually get COVID but it's all in how sick you get. Vaccinated will get less sick.

137 7 93 

 **Lord Fraud #KroenkeOut #ArtetaOut** @Andrew70265613 · Nov 20 · ...
Replies to [@DrEliDavid](#)
Vaccines reduce hospitalisations and deaths.

How can people put Dr next to their name without understanding the absolute basics of science.

The timing of these lockdowns is interesting in light of Paxlovid and the new vaccine mandate. Otherwise, without a new effective drug or a lot more vaccinations on the way, what are you hoping for when you lock down?

Here it's clear what you are hoping for. If the policy works, in two months a lot more Austrians will be vaccinated and a new drug will be available. Spread will come down a lot, and for those who still get infected, death rates will drop dramatically. So rather than kicking the can down the road, this could actually do what a lockdown is supposed to do, and get the corner turned permanently.

On another note and as your periodic reminder, Andrew Cuomo was the worst. His successor Kathy Hochul has the great advantage of not being Andrew Cuomo. In an attempt to make up for that head start, [here she is threatening collective punishment](#) via new restrictions if cases don't drop.

ALBANY, N.Y. (AP) — New Yorkers in upstate communities seeing upticks in COVID-19 cases could again face more pandemic restrictions if infection rates fail to drop, Gov. Kathy Hochul said Tuesday.

The Democrat later [tweeted](#) a note of optimism: saying she'll direct the state health commissioner to relax rules like masks in schools if New York gets “through the holidays without a spike.”

Hochul didn't outline any specific COVID-19 protocols she could put back in place, or how she would define a decline or spike in COVID-19 rates.

But she said she's troubled by vaccine holdouts and signs of rising COVID-19 cases in parts of western and central New York.

“At some point if the numbers don't start on a downward trend, we're going to have to talk about larger protocols,” Hochul said. “I truly hope the community at large will listen to this because it doesn't have to be this way.”

I do give her credit for pretending to be symmetrical, and offering to relax rules like masks in schools if the impossible happens. The ‘no winter spike’ ship has sailed. It still gives hope for the future. Similarly, this is presumably an empty threat along with that empty promise. It is an unspecified threat to impose unspecified restrictions under unspecified conditions, if the people let her down and don’t ‘listen.’ You see, if the same thing happens in New York that happens everywhere else that is similarly cold, *it must be the fault of the people*, who ‘don’t listen.’ Thus they then must be punished.

I am curious what would happen if she made good on her threat. [Buffalo has reinstated a mask mandate.](#)

How much would people tolerate at this point?

Non-Covid Local Aside: Come On Step Inside, You'll Be Taken For a Ride

In non-Covid Hochul news, [here she is](#) bragging that thanks to that sweet, sweet federal cash, we’re going to build out the Second Avenue Subway another 29 blocks for the low, low (estimated! hoped for!) price of \$6 billion dollars, or slightly over \$20 million per city block, or \$2 billion per subway stop that will be used, two of which are new and one of which will be a connection. It’s expected to serve 100,000 riders daily, at the (again, low low low!) cost of \$60,000 per daily rider.

If we reasonably estimate that a daily rider gets marginal value of \$5/ride, that’s \$500k/day in value, or \$182 million/year, so a 3% yearly return on investment. At current borrowing rates, plus a fiscal multiplier, that’s at least reasonable. Could be worse if this is mostly substituting from otherwise available 456 trains, but if this substitution substantially improves conditions on the 456 line, which were usually at capacity in an unpleasant way in non-pandemic rush hours, or this substantially improves traffic uptown and on the FDR, you could potentially generate a *lot* of value. When I was commuting on the 456 I’d have paid a decent amount extra to have better conditions. So looked at this way, the project isn’t a slam dunk at this price, but if you somehow presume it will finish within or near budget it isn’t obviously crazy.

Plus, ‘at the heart’ of it all, this will promote ‘transit equity.’ Quick, someone who understands real estate, solve for the equilibrium.

Except there’s a problem. The population of all of Spanish Harlem is 116,000 people. How are 29 blocks of subway tracks and 2-3 extra stops going to generate 100,000 additional daily riders?

Think of the Children

Speaking of collective punishment, when will our children be able to take their masks off? For a while now, the question has been whether there is an end at all, or there is permanent midnight. Thus, things are so gloomy that almost *any* endpoint, for *any* restriction, provided it will be adhered to, feels like good news. I also try to adhere to the principle of not going after those who take positive action because they didn’t do *enough*, lest we fall prey to the [Copenhagen Interpretation of Ethics](#) and punish both good behavior and action in general.

Despite that, it’s [not like Michael Blume is wrong here](#), there is a limit.



Stop de kindermoord @michaelblume · 5h

...

Outdoor masking, she's talking about outdoor masking when we've known for well over a year that outdoor transmission isn't a thing.

This is blatant collective punishment. The cruelty is the point.



Jackie Goldberg @Jackie4LAkids · Nov 18

#LAUSD News

Starting on January 11, outdoor masking will not be required at @LASchools campuses if at least 85% of the students at that school are fully vaccinated. Indoor mask-wearing will remain in place.

Please get all eligible students vaccinated!

I wouldn't go so far as to say outdoor transmission 'isn't a thing' but having an outdoor mask mandate is still rather ludicrous. And there's no end in sight for the *indoor* mandate that will be in place for most of these children's waking hours. Young children, mostly vaccinated, doesn't matter. Permanent midnight. Holding the release of even the *outdoor* mask mandate hostage via collective punishment to force child vaccinations is exactly what it looks like. And yes, these people don't actively want children to suffer in general, but despite that, no, *exactly because of that*, the cruelty is the point. If you are cruel, it shows you care about being part of the coalition more than you care about not being cruel to children. Otherwise someone might think you cared about not being cruel to children.

FiveThirtyEight's original headline (hint: always look at the URL) is [We Polled Kids About the Pandemic, They're Doing Surprisingly OK](#). They've since noticed that 'surprisingly OK' in context is more a statement about [what we were expecting](#) than it is about how the kids were doing. It's still good news. The kids aren't great, but this was better than my expectations. Maybe they're mostly alright.

They are, however, [still eating lunch outside on the concrete ground](#). Come for the disbelief that kids would do this and the son who 'sits with the kids who have allergies because they get a bench,' stay for the 'it's much safer so I don't see the problem' at the top of the comments.

In Other News

Maybe it's not the lack of heat, [it's the lack of humidity?](#) The theory is that *large drops* in humidity and the resulting adjustment periods cause physical dynamics that make viruses spread, and that's why Covid-19 is 'seasonal' but its patterns look weird. Huge if true! It seems like there are enough other data points that we should be able to confirm or deny the hypothesis. Anyone want to give it a systematic shot?

In Paxlovid news since my post yesterday, a suggestion that Paxlovid won't have much impact [because America can't get a pill to people within three days](#).

But it might be difficult to get the drugs outside a clinical trial setting. Depending on the particular patient, it could involve four individual steps: recognizing symptoms, receiving

a positive Covid-19 test result, being prescribed an antiviral by a doctor, and picking up the pills at a nearby pharmacy.

Each step could prove difficult, Gaffney said, beginning with the challenge of recognizing symptoms during winter, when early signs of Covid-19 might be easily written off as a cold, flu, or allergies. Even if patients do quickly suspect they have Covid, diagnostic tests are still sometimes hard to come by. Many of the patients who test positive won't have primary care physicians. And perhaps worst: The antivirals are ideally taken just three days after symptom onset, meaning the four-step process can't face any setbacks.

Of all the challenges patients will face when seeking the antiviral treatments, the lack of access to efficient testing is by far the largest.

The FDA of course is *also* fully responsible for our lack of testing availability. And it's also responsible for the system that requires the appointment to get the prescription, and the pharmacy to fill it. It's not like this wouldn't entirely be a systemic failure where the government forbids people from access to life saving medicine, it's only a slightly different mechanism.

This isn't three days from *infection*. That's hard. This is three days from *symptoms*. That should be trivial. There exist cheap *rapid* tests that take fifteen minutes.

Thanks to [the murderous madness](#) delaying Paxlovid, we have a month to prepare. There is zero excuse. [FIX IT!](#)

Instead, not the faintest whisper of doing anything but nothing.

I can only imagine a researcher at Pfizer, having now given us first a safe and effective vaccine and now a safe and effective treatment pill, screaming "what does it take? Literally what the f*** do you want from us?" at the top of their lungs at no one in particular.

If we are showing little interest in solving these problems, even now, then are we actually interested in stopping people from dying?

Meanwhile, if you're in need of a bouncer today at Thanksgiving dinner as you demand negative Covid-19 tests, despite the availability of boosters, [maybe it's not about stopping disease?](#)

[CDC data on vaccinations is a contradictory mess, which they were notified about weeks ago](#) and also says on-their-face nonsensical things like a 99.9% vaccination rate among seniors in some areas. I do not think the CDC is making deliberate errors in its data, but any reasonable set of sanity checks would have found this, as would keeping an eye out for people claiming there are errors.

[A detailed analysis](#) claims to show that our social distancing efforts were actively counterproductive and misaimed, reducing activity in all the wrong places and closing enough locations to close to overcrowd those that remained open. I have quibbles with the details, but given that this doesn't include outdoor activities, [it's worse than that](#).

[Eliezer Yudkowsky's current epistemic state](#) with respect to the origins of Covid-19 updated significantly this week, [article here](#). The key questions to ask here are what your prior was before this information (if possible please do pick a number) and then what is the likelihood ratio on this new information? How often does it happen in the worlds where the origins of the virus involved a lab versus didn't involve a lab?

[Richard Hanania notices that he is surprised](#) by the variation in response to Covid-19. I agree that *given other things are as similar as they are* that it was surprising that *the particular responses in question* varied so much. The imposition of mask mandates and shutting down

of life varied greatly from place to place, often but not always in line with partisanship, and this didn't change people's votes much on either side.

I strongly agree that we learned people are more conformist than we thought, that partisanship is stronger than we thought, that extreme government actions can be normalized quickly, and that elected governments on both the state and federal levels have far more rope to do things they feel like doing than we expected, no matter what the laws technically say they're not allowed to do.

Red state America is the outlier where there was real pushback and real skepticism of elites. While that often got in the way of prudent measures, and advocates of such things often were spouting nonsense and being quite annoying about it, this also got in the way of lots of importantly imprudent measures. On net, it's now safe to say that refusing to quietly play along on the margin came out ahead. We will need a healthy amount of such skepticism going forward, if we are to retain some semblance of our freedom and continue believing that we might have rights.

What I also find even more curious than the variation in policy decisions regarding restrictions is the *conformity and uniformity* of other decisions, especially those that prevented us from getting access to information. Things like (and by all means add 'with notably rare exceptions' in places where there were notably rare exceptions):

1. Everywhere, elites agreed what the current Official Facts were, and what was misinformation. Then the facts changed, and they changed everywhere, as did what counted as misinformation.
2. Everywhere we were not allowed to conduct experiments to find out how Covid-19 worked, running only the same handful of experiments that were standard enough to be permitted. We *still* don't know many basic facts.
3. Everywhere we were not allowed to do challenge trials to accelerate vaccine development or test treatments. Everywhere approved the vaccines at roughly the same time with roughly the same level of required evidence, with only minor variations. No one gave even a first look to the nasal vaccine, so it was never tested and we'll never know if it works, and so on.
4. Everywhere treated the FDA and CDC, or similar local authorities that kept to the same official lines, as the arbiters of not only truth but also of what was permitted, and acted as if those decisions everywhere were sane rather than crazy, although there was some variation in how badly various things were botched.
5. Everywhere imposed the same types of restrictions in roughly the same order. Almost no places started 'following the science' and switching up the order as we figured things like 'outdoor transmission mostly isn't a thing.' The biggest variation was to what extent various places continued to place an emphasis on surfaces, 'deep cleaning' and hand washing after they were known to be not important, but that was mostly private action from the start for both logistical and psychological reasons.
6. Everywhere rolled out the vaccine via the same prioritization methods, and mostly stuck to the distributions from the initial studies, although a handful of places instituted first doses first, and some did mix and match faster than others, but that shows how low hanging the fruit had to be in order to be considered at all. No one did fractional dosing.
7. Everywhere used six feet or two meters as the social distance rule despite it being an arbitrary number.
8. Everywhere acted as if the lab leak hypothesis was a conspiracy theory.
9. Everywhere treated variolation (deliberate infection in order to build immunity) as not being a serious proposal. It's now clear it would have been a big win.

Not Covid

Via Marginal Revolution, [a news report](#) that a lot of supermarkets are putting filler stock on their empty shelves to pretend that they aren't out of large numbers of products due to supply chain issues. A question that I haven't seen asked is why they can't fill those shelves with *different* products they don't normally carry, but which might actually sell. In normal times, as I understand it, there are lots of brands that would *love* to get onto supermarket shelves, so much so that they often pay for placement. Surely a deal could be reached here to give those alternatives a shot, and see what happens?

[Itai Sher thread](#) (via MR) about how much deference and trust we owe 'experts,' especially when this is defined by 'I wrote a book about a thing and convinced someone to publish it.' My position is essentially 'some benefit of the doubt on fact questions unless we have reason to be suspicious (in which case trust but verify, followed if needed by don't trust and still verify), but for conclusions or generalizations very little unless you can do a lot better than having written a book or having the right degree or faculty/government position.'

The Lighter Side

[Short twitter thread.](#)

Noah Smith 🐰 ✅ @Noahpinion · Nov 20
At a house party for A.I. people, and saved from having to do a go-around introduction by the timely arrival of a gang of thieves stealing catalytic converters from the cars outside

36 43 591

Noah Smith 🐰 ✅ @Noahpinion · Nov 20
Ultimate San Francisco tweet

6 2 243

Noah Smith 🐰 ✅ @Noahpinion · Nov 20
Friends asked a thief "What are you doing?" He replied "I'm robbing."

The police walked over and chatted with him, but didn't arrest him, and he left.

15 24 270

Noah Smith 🐰 ✅ @Noahpinion · Nov 20
Replying to @Noahpinion
The policeman explained that the guy didn't have the right tools and would not have been successfully able to steal car parts.



John knows nothing @JohnCarltonKing · Nov 20

Please tell me you're kidding

Please

2



2



Noah Smith 🐰 ✅ @Noahpinion · Nov 20

Nope

Who will [stand up to these dastardly muppets?](#)



No idea. It's an ongoing problem. Who will stop the count?

Then again, I still don't really know [what I was expecting](#)?

Los Angeles Times

Los Angeles Times  @latimes · 4h

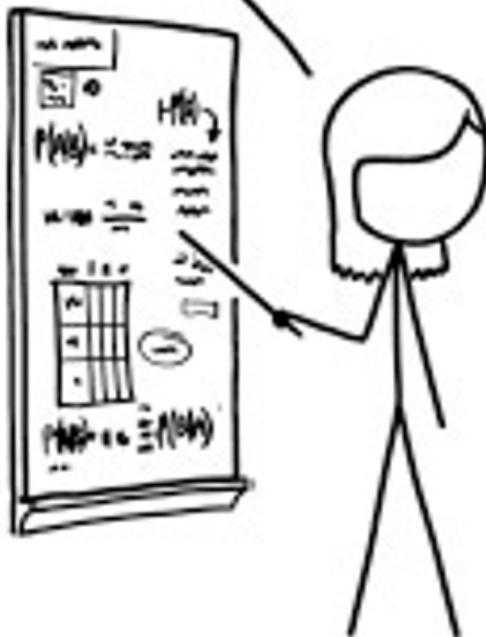
Computer models posted on a California website offer an array of possibilities — both good and bad — for the state's winter.

...

Other times, [I do know what I'm expecting](#).

GIVEN THESE PREVALENCES,
IS IT LIKELY THAT THE TEST
RESULT IS A FALSE POSITIVE?

WELL, THIS CHAPTER IS ON
BAYES' THEOREM, SO YES.



SOMETIMES, IF YOU UNDERSTAND
BAYES' THEOREM WELL ENOUGH,
YOU DON'T NEED IT.

Full credit. This is how this works. This is *exactly* how all of it works.

Why I am no longer driven

(cross-posted from my blog, [Sunday Stopwatch](#))

I used to be very driven. I'm talking "wake up at 4 AM, go run with my dog, take a cold shower, study for two hours, go work, meditate, cook, go to MMA practice, read a book before bed" driven. I was very ready to, you know, get after it. As you might notice from the title, this is no longer true. Why?

The first and obvious explanation is that I'm getting older. I have no way of confirming or denying this - it really could be the case that it's some mashup of biological processes that works independently of whatever my subjective experience of life is. But I'm 29, which doesn't seem old at all, so what are the other reasons?

The first reason I can think of is that normal, everyday life is actually difficult. This may be an entirely personal phenomenon, but I find that I semi-consciously categorize my life into two areas: "takes energy" and "doesn't take energy". Things that take energy are the big important things, like studying for some important professional certification or completing a big project or whatever. And things that don't take energy include all the backend admin work: the cooking, the cleaning, the handling of people, the coordinating, the who's gonna take the dog for a walk, the did you remember that we have a family lunch this weekend, and more.

You don't have to be particularly attentive to realize that the "doesn't take energy" category is completely bogus. It's a scam, one that I've willingly enrolled myself into. My general theory for why I've done this to myself is the implicit attitude that if you do things "right", things from this other category shouldn't take time or energy. Like, these are the basics, right, and the basics shouldn't be hard. I don't know, it doesn't make a lot of sense when I look into it, which is why it's a total scam. Things that don't take energy, take energy.

There's something more that's sorta related. Sometimes, you do a thing that "takes energy" and that's it, you've done it, done. But projects usually require maintenance after you do the project itself, so now you have to, again, do the backend admin work. If you study something, you have to apply it or at least go through some flash cards because otherwise, you're losing it. If you train some skill, same thing. If you get promoted, you have a harder baseline job. If you buy a house, congratulations on your second job!

Then there's also the fact that just a basic, regular day isn't always equally draining. If you have some issue at work, you might be tempted to wave it off as a minor hindrance, a minor annoyance, not something worth your energy and attention as you come back home. Because, hey, it's just this silly little thing, I shouldn't be bothered at all. But you can't talk yourself out of a problem you didn't talk yourself into in the first place. So you come home, and sure enough, you have your regular set of things that you're supposed to do now but which don't take energy, but in addition, you've actually, in reality, had a very, very, very difficult day (and you're pretending you didn't because you feel it shouldn't have affected you that much).

So that's one type of explanation. The other one is disillusionment, or, at least, a change of mindset. I can see a change in the type of YouTube video that I can stomach these days. I used to watch those "How I Keep Productive, Work on 15000 Side Hustles, Run 3 Marathons a Day, and Earn 3 Million Dollars A Month" videos and then I

stopped. I just can't anymore. And in particular, I can no longer watch stuff that takes inspiration out of fiction. I used to have a small theory that went something like this: you are very affected by the circle of people around you, but it doesn't make that much of a difference if the people are fictional or real.

But these days, when I see someone use Goku or Naruto or some other anime character to prove a point, I'm outta there. I may even press dislike. "You see, Goku trained at 10x the Earth's gravity to get himself ready for the confrontation. Inspired by that, I try to push my own boundaries every day." I just can't take it anymore. I mean, it's probably because that was my thing for a couple of years in my early 20s, so I have a too strong reaction right now, but I really think that drawing inspiration from fictional characters is highly, highly overrated.

This is a weird intro to my main point, but I'm talking about a general sense of disillusionment, or, to be more precise, quitting a sprint mindset. Getting inspired by something is very treacherous because you might institute some new rules for yourself, announce to the world the upcoming changes, and then follow through, burning a lot of willpower in the hope that you're constructing a habit. Sometimes it works. Getting into that "get after it" mode, doing a spiritual-energetic-mental sprint, psyching yourself up for that promise of success, embracing *the grind*, however sigma it may be, it's pretty bad for long-term results. I mean, maybe someone can pull it off, but I can't. After hundreds of such offensives, I don't trust them anymore.

I've sorta went the opposite route, and it's a weird mixture of individual components that works for me. I don't think I can write an exhaustive list, but I'll name a couple of things: [coasting in neutral](#), [introducing a Sabbath](#), saying no to people close to me, [following curiosity](#), [writing things down](#)... There's probably more. If this is interesting, I can write about individual things, but it's just stuff I've stolen from smart people.

Anyway, life is now less intense, but more long-term. It's not a sprint, not even a marathon, it's more like a hike. I am no longer driven, in the sense of biting down hard and getting after it all day every day. I rest more, I take walks, I sometimes skip training sessions, and I respect the energy-draining aspects of mundane stuff. I don't ask myself for superhuman performance anymore because it's a short-sighted ask. And I think I get an equal amount done. Most of all, I don't beat myself up anymore for not performing up to some arbitrarily high standard. Life is forgiving, and it's wise to give yourself the spare energy so that you can actually sprint when life isn't forgiving.

Worst Commonsense Concepts?

Perhaps the main tool of rationality is simply to use explicit reasoning where others don't, as Jacob Falcovich [suggests](#):

New York Times reporter Cade Metz interviewed me and other Rationalists mostly about how we were ahead of the curve on COVID and what others can learn from us. I told him that Rationality has a simple message: "*people can use explicit reason to figure things out, but they rarely do*"

However, I also think a big chunk of the value of rationality-as-it-exists-today is in its corrections to common *mistakes* of explicit reasoning. (To be clear, I'm not accusing Jacob of ignoring that.) For example, bayesian probability theory is one explicit theory which helps push a lot of bad explicit reasoning to the side.

The point of this question, however, is not to point to the *good* ways of reasoning. The point here is, rather, to point at *bad* concepts which are in widespread use.

For example:

- **Fact vs opinion.** There are several reasons why this is an awful concept.
 - The common usage suggests that there are "matters of fact" vs "matters of opinion"; eg, I like hummus (opinion) but $1+1=2$ (matter of fact). But common usage also suggests that probabilistic reasoning gives mere opinions, while other modes of reasoning (such as direct observation, and logical reasoning) yield facts. This is inconsistent; it suggests that we can tell whether a belief is an opinion or a fact by examining what it is about (beliefs about subjective things = opinions; beliefs about objective things = facts), while also seeming to need the mode of reasoning by which we arrived at the belief (eg, if I saw a black hole myself, it would be a fact, but if I derived one's existence from unproven physics, it would be opinion).
 - Calling something a fact generally indicates that others are epistemically obligated to believe it. But if it is contentious, then this is precisely what's at issue. So calling something a fact like this is generally useless.
 - We could take "fact" to mean something like "true opinion". But from the inside, this is no different from a strong belief. So again, to call something a fact rather than a strong opinion seems to add no information (whereas, I take it, it's *supposed to* according to common usage).
- **"Purpose" as an inherent property.** In common usage, it makes sense to ask "the purpose of life" because a purpose is a property which lots of objects have. In reality, it only makes sense to think of "purpose" *relative to some agent*, as in "I made this for this purpose". Common usage allows purpose to be agent-independent because there are lots of things (tables, chairs, silverware, etc) which have purposes largely independent of agent (most people use tables to set things on for convenient reach, chairs to sit on, silverware to eat, etc). However, in cases which aren't like that, the language doesn't make sense without explanation (but people treat it like it does).

These are intended to be the sort of thing which people use unthinkingly -- IE, not popular beliefs like astrology. While astrology has some pretty bad concepts, it is explicitly bundled as a belief package which people consider believing/disbelieving. Very few people have mental categories like "fact-ist" for someone who believes in a

fact/opinion divide. It's therefore useful to make explicit belief-bundles for these things, so that we can realize when we are choosing whether to use that belief-bundle.

My hope is that when you encounter a pretty bad (but common) concept out there in the wild, you'll think to return here and add it to the list as a new answer. (IE, as with all LW Questions, I hope this can become a timeless list, rather than just something people interact with once when it is on the front page.)

Properly [dissolving](#) the concept by explaining why people (mis)use it is encouraged, but not required for an entry.

Feel free to critique entries in the comments (and critique my above two proposals in the comments to this post), but as a contributor, don't stress out about responding to critiques (particularly if stressing about this makes you not post suggestions -- the voting should keep the worst ones at the top, so don't worry about submitting concepts that aren't literally the worst!).

Ideally, this would become a useful resource for beginners to come and get de-confused about some of the most common confusions.

What would we do if alignment were futile?

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

Eliezer's [recent discussion](#) on AGI alignment is not optimistic.

I consider the present gameboard to look incredibly grim... We can hope there's a miracle that violates some aspect of my background model, and we can try to prepare for that unknown miracle

For this post, instead of debating Eliezer's model, I want to pretend it's true. Let's imagine we've all seen satisfactory evidence for the following:

1. AGI is likely to be developed soon*
2. Alignment is a Hard Problem. Current research is nowhere close to solving it, and this is unlikely to change by the time AGI is developed
3. Therefore, when AGI is first developed, it will only be possible to build misaligned AGI. We are heading for [catastrophe](#)

How we might respond

I don't think this is an unsolvable problem. In this scenario, there are two ways to avoid catastrophe: massively increase the pace of alignment research, and delay the deployment of AGI.

Massively increase the pace of alignment research via 20x more money

I wouldn't rely solely on this option. [Lots](#) of [brilliant](#) and [well-funded people](#) are already trying really hard! But I bet we can make up some time here. Let me pull some numbers out of my arse:

- \$100M per year is spent per year on alignment research worldwide (this is a guess, I don't know the actual number)
- Our rate of research progress is proportional to the square root of our spending. That is, to double progress, you need to spend 4x as much**

Suppose we spent \$2B a year. This would let us accomplish in 5 years what would otherwise have taken 22 years.

\$2B a year isn't realistic today, but it's realistic in this scenario, where we've seen persuasive evidence Eliezer's model is true. If AI safety is the critical path for humanity's survival, I bet a skilled fundraiser can make it happen

Of course, skillfully administering the funds is its own issue...

Slow down AGI development

The problem, as I understand it:

- Lots of groups, like DeepMind, OpenAI, Huawei, and the People's Liberation Army, are trying to build powerful AI systems
- No one is very far ahead. For a number of reasons, it's likely to stay that way
 - We all have access to roughly the same computing power, within an OOM
 - We're all seeing the same events unfold in the real world, leading us to similar insights
 - Knowledge tends to proliferate among researchers. This is in part a natural tendency of academic work, and in part a deliberate effort by OpenAI
- When one group achieves the capability to deploy AGI, the others will not be far behind
- When one group achieves the capability to deploy AGI, they will have powerful incentives to deploy it. AGI is really cool, will make a lot of money, and the first to deploy it successfully might be able to impose their values on the entire world
- Even if they don't deploy it, the next group still might. If even one chooses to deploy, a permanent catastrophe strikes

What can we do about this?

1. Persuade OpenAI

First, let's try the low hanging fruit. OpenAI seems to be full of smart people who want to do the right thing. If Eliezer's position is true, then I bet some high status rationalist-adjacent figures could be persuaded. In turn, I bet these folks could get a fair listen from Sam Altman/Elon Musk/Ilya Sutskever.

Maybe they'll change their mind. Or maybe Eliezer will change his own mind.

2. Persuade US Government to impose stronger Export Controls

Second, US export controls can buy time by slowing down the whole field. They'd also make it harder to share your research, so the leading team accumulates a bigger lead. They're easy to impose: it's a regulatory move, so an act of Congress isn't required. There are already export controls on narrow areas of AI, like automated imagery analysis. We could impose export controls on areas likely to contribute to AGI and encourage other countries to follow suit.

3. Persuade leading researchers not to deploy misaligned AI

Third, if the groups deploying AGI genuinely believed it would destroy the world, they wouldn't deploy it. I bet a lot of them are persuadable in the next 2 to 50 years.

4. Use public opinion to slow down AGI research

Fourth, public opinion is a dangerous instrument. It'd make a lot of folks miserable, to give AGI the same political prominence (and epistemic habits) as climate change research. But I bet it could delay AGI by quite a lot.

5. US commits to using the full range of diplomatic, economic, and military action against those who violate AGI research norms

Fifth, the US has a massive array of policy options for nuclear nonproliferation. These range from sanctions (like the ones crippling Iran's economy) to war. Right now, these aren't an option for AGI, because the foreign policy community doesn't understand the

threat of misaligned AGI. If we communicate clearly and in their language, we could help them understand.

What now?

I don't know whether the grim model in Eliezer's interview is true or not. I think it's really important to find out.

If it's false (alignment efforts are likely to work), then we need to know that. Crying wolf does a lot of harm, and most of the interventions I can think of are costly and/or destructive.

But if it's true (current alignment efforts are doomed), we need to know that in a legible way. That is, it needs to be as easy as possible for smart people outside the community to verify the reasoning.

*Eliezer says his timeline is "short," but I can't find specific figures. Nate Soares gives a very substantial chance of 2 to 20 years and is 85% confident we'll see AGI by 2070

**Wild guess, loosely based on [Price's Law](#). I think this works as long as we're nowhere close to exhausting the pool of smart/motivated/creative people who can contribute

What exactly is GPT-3's base objective?

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

[Probably a noob question]

I'm thinking about what an inner alignment failure might look like for GPT-3. This would have to involve some deployment context in which GPT-3 performs significantly worse (by the standards of the base objective) than it did in training. (It would involve other things too, such as GPT-3 being a mesa-optimizer.)

But to say how well GPT-3 performs on some prompt not in the training dataset, we have to have a definition of the base objective that extends beyond the training dataset. If the base objective only makes sense in the context of the training dataset, then inner alignment failure is impossible by definition.

Is the base objective "Predict the next word?" Or is it "Predict the next word, supposing what you are reading is typical 2019 Internet text?" Or is it "Predict the next word, supposing what you are reading is a random-with-the-following-weights sample from dataset D? [where D is the dataset used to train GPT-3]" The third option is in some sense the best, because it most closely fits what we actually did to train GPT-3. But note that the logical extension of this line of reasoning is to prefer a fourth option: "Predict the next word, supposing what you are reading is a random-with-the-following-weights sample from dataset D' [where D' is like D except that it doesn't contain any of the bits of text that GPT-3 happened to not see in training, and the randomness weights are chosen to more accurately yield the data points that GPT-3 in fact saw]."

The problem with these last two answers is that they make it *undefined* how well GPT-3 performs on the base objective on any prompt that wasn't in D, which then rules out psuedo-alignment by definition.

From the Risks from Learned Optimization paper:

In such a case, we will use *base objective* to refer to whatever criterion the base optimizer was using to select between different possible systems and *mesa-objective* to refer to whatever criterion the mesa-optimizer is using to select between different possible outputs. In reinforcement learning (RL), for example, the base objective is generally the expected return. Because the mesa-objective is not specified by the programmers, mesa-optimization opens up the possibility of a mismatch between the base and mesa-objectives, wherein the mesa-objective might seem to perform well on the training environment but lead to bad performance off the training environment. We will refer to this case as pseudo-alignment below.

Expected return in a particular environment/distribution? Or not? If not, then you may be in a deployment context where you aren't updating the weights anymore and so there is no expected return, or at least it's close to 0 because there's only any return if you can convince people to start updating your weights again!

I worry I am just confused about all this. Hence why I'm asking. **What is GPT-3's base objective?**

Investigating Fabrication

Note: This essay was written in response to “[Lies, Damn Lies, and Fabricated Options](#)”. I expect it’ll be pretty confusing to read this one without having read that one first. However, I’m not the boss of you, and it isn’t obvious to me anyway which reading order is best overall. If prolonged confusion makes you grumpy, I recommend reading Duncan’s essay first. But if you try it backwards, I’d love to hear how that goes.

1. Motivation

When I read Duncan's essay "[Lies, Damn Lies, and Fabricated Options](#)," it seemed to me to point toward something that's worth knowing for real, rather than just worth knowing *about*.

I think it does a good job of helping the reader build a new concept that could conceivably be taken as a hypothesis in important situations. The hypothesis is, "Perhaps this thing that seems to me like an option isn't *really* a possibility, and it only seems like one because I'm in the grips of a map/territory conflation." (Or, at least, "Perhaps the person I'm arguing with thinks their made-up option is real because [etc.]".)

But I wasn't satisfied with my new concept "fabricated options". There were some super important questions that I'd need to answer before I could put it to use.

For instance,

- Which situations are the crucial ones for promoting this hypothesis to attention?
- How do I actually move my mind to recognize that the option I'm considering might be fabricated?
- Once I've posed the hypothesis, how do I confirm or deny it?
- When I've successfully identified a fabricated option, how can I prevent the usual harm?
- And, my favorite: How can I stop making this kind of mistake in the first place?

So I set out to investigate option fabrication (the act), and what my brain is up to in the moments just before a fabricated option appears. I hoped that if I could learn where fabricated options come from, and what they are made of, I'd be a long way toward answering many of the practical questions on my list.

2. Approach

I did [naturalism](#) to this. (One of the reasons this essay felt worth writing and sharing is to give people more chances to see what naturalism looks like in practice.)

I assumed that there's sometimes A Thing Going On in my brain that in some ways resembles what Duncan called "fabricated options", and that my default understanding of the Thing is worse than I'd prefer. I set out to observe the Thing myself, directly and in real time, and to gradually build my own understanding of

whatever I found. I tried to observe the Thing both in the "field" (as it occurs naturally in the course of daily life) and in the "lab" (as it occurs when I deliberately create controllable conditions that might give rise to it).

This was not a full naturalist study (which usually takes at least a month), but was instead a "let's see what comes of a week or two" dalliance. Still, it turned out to follow the arc of a mostly complete study: Sensitization, zooming in, zooming out, and experimentation.

3. Lab Work, Part 1

In accordance with the “try it immediately” heuristic, I started out with a relatively lab-like exercise right after reading the essay. I asked myself, “What is it like to fabricate options?” and in a search for relevant data, I began to scan my recent memories for “options” I might have “fabricated”.

Multiple topics came to mind right away. One of them was “top surgery”, which I was scheduled to have in just a few days.

A thought occurred to me; it was one I’ve had before (it felt familiar as it arose), though it wasn’t one I’d examined closely, or tried to put into words. I tried this time, though, with extra attention:

“Here’s what I will do. I will have top surgery, but I’ll skip all those awful parts with pain and anesthesia drugs and so forth. I’ll just push a button, and my boobs will be gone. Simple.”

Although this was pretty clearly an instance of a “fabricated option”, I was at this point less interested in the thought itself, and more interested in what led me to turn my attention toward “top surgery” in the first place. What caused me to (correctly) expect to find The Thing there? I wanted to know in what directions I should open my introspective sense if I want to pick up more information about fabricated options. What properties does “top surgery” have that mark it as the natural habitat of the “fabricated options” creature I would like to study?

The “boob elimination button” thought was one I seemed to recreate frequently as I considered “top surgery”, so I had several opportunities to observe the conditions that gave rise to it. Here’s what always preceded it:

When I thought about top surgery, I felt good things, and I also felt bad things. As I imagined having a masculine chest and never having breasts again, I felt good. As I imagined going through surgery and recovering from it, I felt bad. And when my attention fell on the fact that I had *scheduled* top surgery, that it was in my future by default, I began to do and feel something I call “trying to squirm out from under the problem”.

“Trying to squirm out from under the problem” felt like this:

When I thought about top surgery, it was like finding that something hard and heavy was on top of me. I did not like being under the hard heavy thing, and I didn’t know how to lift it. I was getting squished. It was uncomfortable under there, so I squirmed: I pushed up on this part of it with this hand, and that part of it with that foot, and I wriggled and writhed and tried to work myself out sideways.

That’s what it *felt* like, in my body and my emotions.

The *thoughts* that happened while I was squirming went like this:

Top surgery is scheduled? Ugh, I don’t like it.

Maybe I could cancel top surgery? No, that doesn't work, because then I'd keep having boobs.

How about I go through with top surgery then? No, that's no good either, because then I'd have to experience surgery and recovery.

Well, let's see... Maybe... Maybe I could cancel top surgery?

sigh Still no, for the same reasons as last time.

I watched this pattern with multiple other topics, too, and it went almost exactly the same way with each of them, always with the same feelings of being trapped, squirming, and repeatedly hitting walls. The general form:

Well, [problem]. I don't like it.

I guess I could [option]? No, because [cost].

How about [second option]? No not that either.

Well maybe [the first thing again]? Still no.

I go through this cycle a few times, and as I do, it's as though I'm banging my head, elbows, and knees against the walls of a much-too-small space. Inevitably, though, I at some point have a thought that goes,

What if [good parts of first option, and good parts of second option], but not [bad parts of either option]?

In the moments leading up to this new kind of thought, I usually feel frustration, restlessness, and desperation. There's a bright flash of a moment that I want to label, "FINE!", as shouted by someone stamping their foot or slamming a door. Then there is a "grasping" and a "stitching", as I weave together "just the good things" and "without the bad ones".

Finally, there is a feeling like a dungeon wall falling away to reveal a sunny meadow, which, if I watch closely enough, has some kind of stuttering dream-like quality to it. It's as though I had to fall asleep, or trip through a magical door, or slip somehow, before I could experience it. But the dreaminess is quiet, and it's largely drowned out by relief (if only briefly).

Another thought I found by searching for a prolonged squirming feeling was, "I know! What I will do is, I will experience winter, with snow and Christmas and all of that, and instead of getting depressed like usual, this time I just won't get depressed!"

4. Lab Work, Part 2

The other thing I did right after reading Duncan's essay was order a gyroscope. Three days later, I played with it.

The main question I was investigating was, "What exactly is it like to imagine a gyroscope that's not weird?"

This turned out to be a bit of a wrong question. Not wrong in the sense that I should have focused my attention on a different question instead; but wrong in the sense that I was employing some nonsensical concepts while asking it.

I tried very carefully to "imagine a gyroscope that's not weird". And what I found is that I'm not capable of imagining a gyroscope. Or an apple, or anything else. Not in the way I naively tend to think.

What I *can* do is conjure a blurry round-ish image and label it "spinning disk". I can conjure a crisp, bright tactile sensation of a sharp pressure on my finger and call it "where the gyroscope balances". I can conjure the sensations I'd expect to feel while tilting my wrist as my hand grasps a gyroscope, and I can imagine the pressure changing just as it would if I were holding a ball, or a phone, or a non-spinning gyroscope.

And if I'm not very slow and careful, it turns out, I can conclude from this imagining that some gyroscopes behave like solid objects even while their disks are spinning. Because hey, I've got a gyroscope in my imagination right now, and it's not being weird at all!

Only, I *don't* have a gyroscope in my imagination.

It turns out that my brain is confused about imagination.

It conjures visual images and textures and stuff, things that resemble "a gyroscope" or "an apple" or "the way my mother drives a car". It conjures them in order to manipulate them, to see what would happen under various circumstances. "What if I threw an apple at the wall?" I imagine doing so, and the thing I'm imagining goes thud-smoosh as it bounces off the other thing I'm imagining. This is most of how I predict stuff.

The problem is, my brain sort of thinks that when it manipulates a shiny red fruit-smelling object in imagination, what it's manipulating *is an apple*.

What it's really manipulating is part of itself, its own expectations and associations and so forth. When I use my own expectations to answer a question like, "What if I threw an apple at the wall?", *of course* the answer I get from imagination is whatever I already expected.

If my expectations are wrong, how will my imagination discover my error? It might let me discover my error by showing me that I have two conflicting expectations, which could give me space to ask whether one of them is wronger than the other (or something). Which is why it's often worthwhile to deliberately investigate my expectations. This is (part of) what's going on when people change their mind when asked to bet, or when they do CFAR-style Murphyjitsu.

But imagination won't let me check a single expectation by itself, because it's taking that expectation as input. It's expectation working with expectation. My imagination is going to play along with itself. Even if I expect to notice some kind of contradiction, and cause myself to imagine that I've noticed the expectation being broken in some way, that brokenness isn't necessarily going to correspond to anything in the territory. I can tell my brain to see how a fantasy is a fantasy, and it's going to succeed, but that success will be another fantasy.

Reality, by contrast, resists individual mistaken expectations, and it does so lawfully.

Noticing the total lack of any feeling of *resistance* is what turned out to be so important about this session with the gyroscope. Whatever thing I was doing as I attempted to "imagine a gyroscope, but not weird" was full of ease, at least after I skipped over the question of what exactly the thing in my imagination was made of. The experience was full of effortlessness. It was full of "of course" and "simple" and "just".

And the particular flavor of ease I found here was similar to the "open meadow" from before, the way I could *slide* into planning for "just the good things, without the bad". It's like shifting a car into neutral. There's a tiny skip, and then all the resistance is gone.

5. Field Work, Part 1

By "field work," what I mean is that rather than engineering experiences specifically for the purpose of trying to investigate The Thing ("fabricated options," or whatever it is), I instead went about my day hoping to run into The Thing in a more natural, unplanned way. I hoped to encounter the phenomenon in its natural habitat, so to speak.

Based on my early lab work, I made some guesses about what I might see or feel if my brain was just about to start doing The Thing, or had just done The Thing, or if I might be near The Thing in some other way. Whenever I ran into something that made me think I might benefit from paying extra attention, I snapped my fingers and took a moment to check what was going on in my thoughts and in my environment. Sometimes, I made a note in my phone to record my observations for later review.

I snapped my fingers many times. Each time, I updated a little bit about how I would know, next time, that *this* is among the experiences I'm watching for. I started with the squirming-out-from-under-a-problem feeling, but the longer I watched myself thinking and feeling in daily life, the more a sensation of *discontinuity in imagination* seemed like the bright shining landmark indicating that "the important thing about fabricated options lives here".

"Discontinuity in imagination" wore a few different phenomenological guises.

One was "sliding", like slipping down a wet hill over a patch of algae. I'd be imagining something, then I'd imagine the next part of the thing, and in between I did not so much "concretely imagine" as "slide" or "slip over".

Sometimes it felt like "skipping" or "tripping", catching my shoe on a root and falling blankly for a moment before resuming my gait as before.

Sometimes it felt like a hollowness, or an empty space. In fact, in most cases where I took the time to watch closely after snapping my fingers, I discovered an emptiness behind some kind of superficial imagining. Something like a facade.

I happen to have recorded my thoughts in a log entry as I observed an experience that involved both hollow-facade-ness and sliding. Here's an excerpt that goes into detail on that experience, with apologies for its length; this is optimized for me talking to myself more than for being nice to readers.

i can easily ask myself, "what's the fairy tale fantasy world thing you want here?" and it's an obvious question, a question that completely fits the feeling of this mental space, and the answer is also obvious: i want to stay up late relaxing and fucking around and acting like i have all the time in the world and not trying to do anything in particular at all, and i also want to wake up in the morning refreshed awake alive alert enthusiastic. what i want is for there to be an additional six hours tonight, just inserted between the usual hours of 22:00 and 23:00.

there's a click when i describe the extra hours. there's even a flow or a gravity toward filling in the details. the clock would say "22:59", then it would say "22:60", then "22:61", and it would just keep going like that until "22:419".

the filling in of details responds somewhat to little concerns i have, hints that something is wrong: after imagining "22:419", i feel a tug, and when i follow the tug i think about "other people". i try out a couple of things. "maybe they're all asleep and they don't notice?" but no, many of them are still awake. i try again: "a time bubble. there's a bubble around just my room where six extra hours happen tonight." and as i think that, there's a relaxation, and then a part of me that's watching all of this is laughing as though at a puppy trying to carry a too-long stick through a narrow doorway.

it's very interesting, though, this relaxation after "time bubble". it's actually a space bubble, right? it's a bubble of space in which time is different, whatever that might mean. my calling it a time bubble highlights how not-even-a-concept this is. it's like an Old West movie set with just the fronts of the buildings and nothing behind them. i can't describe the architecture of the time bubble because it's not even a building.

yet the facade is convincing enough that some part of me reacts as though it's arrived in town and now it can drink the last of the water in its canteen because there's obviously a saloon right there on the corner. "it's ok, there's a time bubble. let's move on to the next question."

what's that like from the inside? it's like painting, and also like pong. a concern hurtles toward the fantasy of "six extra hours", and in response to it i throw some paint up on just the right part of the canvas to block it. there aren't any details yet, but that's ok, because i only need to block the concern on the surface, i only need to convince it to turn away. "what about other people? time bubble!" and if the concern doesn't immediately turn away, that's also all right, because i have no shortage of paint. the more time i spend thinking about it, the more detail i can imagine.

and perhaps the most interesting thing, here, is the way that this sort of imagining feels like discovery. answers like this come from the depths of the association network, from the source of lateral thinking and mnemonic skill. it's not slow deliberate reasoned thought whose pieces you can easily see. the feeling of discovery is accurate, in that i'm discovering how my mind is shaped.

but if i watch closely, there's a sliding. i feel the "discovery", and i sneakily, slippingly infer that i'm discovering the shape of the external world. i slip-slidably infer that i'm learning about how time bubbles work.

discovering the external world feels different. knowing what feels different in discovering the external world vs discovering the shape of your own association network may be the way to recognize that you are painting in response to the pong balls, rather than noticing when the pong balls bounce off of pre-existing walls.

painting—or, perhaps i should say "fabrication", because this i think is the precise cognitive activity at the heart of "fabricated options"—has a characteristic *lightness*. there is no inertia. there is no resistance, no need for exertion or endurance. there is only dream-like ease. it's like gymnastics in super low gravity. it's like Minecraft in creative mode. it's like casting "accio solution" in the Harry Potter universe and watching things fly across the room right into your hands.

6. Field Work, Part 2

Here are some other times when I snapped my fingers.

How about I eat the chocolate chip cookie but not the calories?

This one's classic. I tried to combine the good parts of two options while leaving out the bad. The fabrication revealed itself most memorably in the way it felt to answer my own probing questions:

How will I avoid eating the calories?

They'll leak out first.

How will I cause this leaking?

I'll break the cookie in half.

What do calories look like as they leak out?

Little golden musical notes.

This had the easy, automatic quality of facade-painting I'd identified before. "Accio solution!"

My toes should not touch each other.

I caught a hint of "confused imagining" while wearing regular socks instead of toe socks. I usually wear toe socks because I don't like my toes touching each other.

It took me a while to figure out why "my toes touching" caused me to snap my fingers (I'm deliberately loose/generous with the finger-snapping rather than limiting myself to finger-snaps I already know I can justify), but I did eventually zero in on it. I was trying to force my toes not to stick together by imagining (incessantly) that although they touched, the skin experienced no friction. ("See, world? Here's an existence proof, right here in my imagination. Update!")

There was a blankness and a tripping-over feeling between "imagining toes touching" and "imagining no friction between them". If you'd asked me, "How is there no friction?" I wouldn't have been able to immediately respond with something like "the skin lacks texture". I was skipping over that part. I would only have given you, "I don't know how there's no friction, there just isn't!"

It's not a central example of a "fabricated option", but it's the same mental patterns played out at a pretty low sensory level. I suspect that many instances of "frustrated should" involve something like this.

"I'll just be fine."

This was the first time that what I've so far learned from this study seemed to assert itself and cause a substantially different outcome from the default.

I was in San Francisco on my way to a doctor's appointment. There were a lot of sounds. The parking garage elevator said "Eeerrrr! Eeerrrr! Eeerrrrr!" in a terribly grating voice. The cars whooshed and honked. The crowds of strangers shouted to each other. It was all horribly chaotic and unfamiliar. I could feel myself closing off internally, contracting as though preparing to huddle down into a ball on the floor in some concrete corner.

But I had a mission. Not only did I want to make it to the appointment, but I wanted to talk with the doctor, to ask him questions and to understand what he said.

There was a moment of gathering forces. A moment of forcefully moving the outer shell of my mind into a strong protective shape. And with the gathering, I started to say to myself, as though declaring or proclaiming, "I'll just be fine." I began to decide that instead of being overwhelmed, I would "just" "be fine".

I say "began", because I didn't get very far.

I noticed something that indicated an instance of this thing I've been studying. It wasn't the word "just" that stood out, even though that could have tipped me off.

Instead, it was the discontinuity. There was a feeling like tripping over uneven flagstones. It came from the way the *outer shell* of my mind had shifted, without more central parts having followed. It was a feeling that reminded me of "pulling one over on myself".

I paused, to pay attention to what was happening.

I checked whether I really believed I could "just be fine", and the answer was "sort of". I knew that I definitely *could* do something with my mind that would allow me to walk through the city, and get all the way through my appointment, while barely even noticing my own distress, let alone exhibiting it outwardly. I could do something that would allow me to cross the street safely, to ask the questions that mattered, and to understand the responses.

The facade I'd begun to paint, though, left out a crucial component of this "deciding to be fine": the price. The neurotypical mask is heavy, and there is always a price for wearing it. I pay in exhaustion and future dysregulation.

I did actually decide to pay that price, in this case. But the decision I ended up making was "pay the price for 'being fine'". "Just be fine" was never a real option.

7. No Conclusion; Current Direction

The question I was left with, after the experience in San Francisco, was "Why did I try to deceive myself like this in the first place?"

If I hadn't caught myself in the middle of fabricating an option, I would still have paid the price for "being fine" in downtown San Francisco. I would even have paid a higher price than necessary, because I would have been relatively profligate with my

masking, instead of having turned my conscious attention to the tradeoffs. So what was the point of fabrication?

I'm not sure yet, but I have a pretty confident guess.

I was already terribly overwhelmed when this happened, and I was desperate for a way out. Desperate enough that I was willing to hallucinate a temporary escape. The hallucination couldn't actually cause me to escape the situation, but it could let *that one desperate time slice* experience a piece of relief.

My main conclusion so far from my dalliance with fabricated options is that fabrication is a response to feeling trapped. That's where it comes from, and what it's made of. It's a strategy for getting out of a too-tight possibility space, and its downfall is the way it moves only in imagination without staying in touch with the real world.

I suspect that if I'm going to get things right around "fabricated options" in the long run, I will need a handful of alternative strategies to directly address the sources of the distress. My brain engages in fabrication for a reason. If I want it to do something different, I'll need to address that reason in some other way. The causal need isn't just going to stop being a causal need.

I would bet that the most successful alternative strategies will be space-making. Things like:

- bother to fully notice that I'm uncomfortable
- take three breaths
- find one way to reduce input
- do one fewer thing at a time
- reprioritize
- name one simple innocuous thing nobody can stop me from doing, then do it (e.g. "spin around")
- ask for help
- lower the stakes

If I find myself making up fake options to escape the bind I seem to be in, I think it's likely I've focused so much of my attention on that one particular bind that I'm no longer aware of the ways I'm *not* trapped. If I make myself more aware of the spaces through which I *am* free to move, I might find I'm less desperate to hallucinate an escape, and more comfortable with patiently investigating the true shape of the world.

Appendix

Chris Voss negotiation MasterClass: review

This post is about the [negotiation MasterClass taught by Chris Voss](#) as well as a related [MasterClass Session by Chris Voss](#) and more broadly about the techniques and worldviews shared in those. I [watched the original MasterClass in April 2020 and watched the Session in December 2021](#).

My post discusses some of the relationship -- similarities and differences -- with rationalist advice. These portions should be of particular relevance to LessWrong and are part of the reason for posting to LessWrong.

This is a fairly lengthy post, so please use the sections to navigate to the portions you are most interested in.

I cover a lot in this review:

- What is this a review of?
- Why am I doing this review?
- Worldview
- Negotiation principles
- Negotiation techniques
- Other ideas
- General concerns

ETA 2021-11-24: I added a few more subsections and made some edits to existing content. You can see the change set [here](#).

ETA 2021-11-25: I added a subsection in response to a point raised in the comments. You can see the change set [here](#).

ETA 2022-01-02: I expanded and restructured the review to cover the MasterClass Session that I consumed in December 2021. The original post, written in November 2021, did not cover the Session. You can see the change set [here](#).

What is this a review of?

Who is Chris Voss?

[Christopher Voss](#) worked at the FBI as a lead hostage negotiator. In 2008 he left to form his own company, [The Black Swan Group](#), that offers coaching to people on negotiation techniques. Their customers include people in real estate and many other kinds of business.

What is MasterClass?

[MasterClass](#) is a player in the growing edutainment space, combining education and entertainment. There are 100+ "classes" in MasterClass. Each class is taught by a subject matter expert (the teacher), and includes several videos (ranging from 2 to 30

minutes, usually 5 to 20 minutes). The typical video format just includes the teacher speaking to the camera, but there are occasionally interactive sessions with other participants, or other co-instructors.

Course notes can be downloaded. There are also community features on MasterClass.

You may have seen video ads for MasterClass on YouTube!

What is MasterClass Sessions?

In late 2021, MasterClass introduced [Sessions](#), a more interactive form of MasterClass that is designed a bit more like a course. A few differences between traditional MasterClasses and sessions:

- There are stronger nudges to consume a session chronologically; later activities in a session only unlock after you mark earlier activities as completed.
- While most of the activities in both regular MasterClass and sessions are video consumption, the sessions tend to feature more videos where you, the viewer, is expected to do something active, such as observing or counting something. Although so far these videos don't have interactivity at a technological level (e.g., you are not supposed to click when/where something happens), a conscientious viewer can get the benefit of active learning.
- The sessions also include explicit prompts for viewers to enter information on their progress or thoughts. You have to enter something to proceed (you can't just skip it) though if you really just want to proceed, you can enter gibberish and check the box to not share it with others.
- As far as I can make out, unlike regular MasterClasses, you cannot buy sessions individually; they're only available as part of a subscription.

What does this cost?

A MasterClass annual subscription all-access pass costs \$180; this would allow you to consume both the regular MasterClass and the MasterClass Session within the year you got the subscription for.

A single MasterClass (such as Chris Voss's) can be purchased for \$90. As far as I can make out, sessions cannot be purchased individually, and you need to have a subscription.

How long is the regular Chris Voss MasterClass?

There are 18 videos and the total video time is 3 hours 4 minutes.

How long is the Chris Voss MasterClass Session?

MasterClass doesn't include a total length, and it was tedious to add up the exact lengths, so I just added up the rounded lengths available on the [session page](#). The total that I could count came to 5 hours 50 minutes.

What aspects of the regular Chris Voss MasterClass does my review cover?

The MasterClass includes three live exercises, that I do *not* discuss here (since they are more for demonstration purposes of a range of techniques). I also do not discuss the historical hostage negotiation examples that Chris Voss discusses (a bank robbery, the [Jill Carroll kidnapping case](#), and the [Dwight Watson \(tobacco farmer\) case](#)). I also skip some other material, including discussion of bargaining (using the [Ackerman model](#)), that is outside of the main negotiation techniques.

The class guide is 14 pages long, including a cover page (this blog post overlaps quite a bit with the class guide).

What aspects of the Chris Voss MasterClass session does my review cover?

By and large, I do not discuss the session separately; I mostly only mention ideas in the session that are *not* in the regular MasterClass. Also, I consumed the session and added information about it well after writing the original post, and have relatively limited time.

The session includes a lot of demonstrations involving Voss's colleagues at The Black Swan Group (Brandon Voss, Derek Gaunt, and Sandy Hein). I don't discuss any of the demonstrations individually (though I found them very useful to watch!) and I mainly cover only the insights from the.

Are there other ways of watching the MasterClass without paying the subscription fee?

You can sign up and cancel within 30 days to get a full refund (however, you should check if this option is available when you sign up, as MasterClass may change its policies on this front). This gives you enough time to consume the Chris Voss MasterClass.

You may be able to sign up for a free week; however, I'm not sure what subset of the videos in the MasterClass are available. More details [here](#).

Are there other ways of accessing the information without watching the MasterClass?

Many of the techniques discussed in the MasterClass are discussed online, including in videos from Chris Voss and others. I have more links for each of the individual techniques, so you can use my post to basically get a "free" version of the MasterClass. You can also Google around for more videos and written materials on the techniques.

Voss has also co-authored a book *Never Split The Difference* that covers these techniques. It's [available on Amazon](#); at \$11.99 for the Kindle edition it is considerably cheaper than the MasterClass. There is a [Reddit thread](#) comparing the book and the MasterClass and consuming both.

[Masterwiki.how](#) can be a helpful free version of MasterClass, but it does not seem to have Chris Voss's MasterClass.

What's the value-add of the MasterClass beyond the free material in and linked from this blog post?

My blog post didn't exist at the time I bought the MasterClass; I think in the absence of such an alternative, the MasterClass was worth paying something for. In fact, I would say that the exposure to Chris Voss's ideas was worth the price of two full years of MasterClass subscription for me, though I could have cancelled after 30 days once I had consumed the Chris Voss MasterClass.

Now that this blog post exists, and given the great amount of online material, the value-add of the MasterClass is less clear. I think there is still a case for it, but it probably isn't worth spending a full year's subscription of \$180 just for this class, nor is it worth spending \$90 to buy it. With that said, if you find several other MasterClasses that you consider worth watching, the Chris Voss MasterClass could tip the scales.

Given the ability to cancel after 30 days, it seems worth trying it out if you think there's a chance it will be worthwhile.

Since rationalists have often spent \$1,000+ on CFAR workshops and found it worthwhile, I think there's a good chance that many will find the class worth paying for, even after having access to the material in the blog post.

Are there other worthwhile MasterClasses?

I found [Daniel Pink's sales and persuasion MasterClass](#) to be similar and relevant. There are a lot of similarities and some differences between Pink's and Voss's advice on the topics where they overlap. [Here](#) is a good review of Pink's MasterClass as well as some discussion of the similarities and differences with Voss's.

In general I have found MasterClass to be worth the subscription cost for at least two years, but much of it might have to do with the specific topics I am interested in. You can check [here](#) what MasterClasses I have watched.

What are some interactive discussions where these techniques are critically examined?

The MasterClass is mostly just Chris Voss speaking, with nobody challenging him. The session does a better job of showcasing interaction among multiple people, but they're still all people at The Black Swan Group, so it doesn't necessarily show as much as possible how their ideas could be challenged or questioned by outsiders.

I like to see people questioned about their ideas, ideally with question that I have or would have if I thought more about it. Since watching the original MasterClass, I've [watched several interviews of Chris Voss](#); I list below the ones I found most interesting:

- [Master the Art of Negotiating in Business and Life: Lewis Howes](#)
- [#MasterClassLive with Chris Voss](#)

- [The Knowledge Project #27 — Chris Voss](#)
- [WHY SUCCESS Comes From Mastering Negotiation In BUSINESS & LIFE | Chris Voss & Lewis Howes](#)

Why am I doing this review?

I think the MasterClass and the material it covered were pretty good

Chris Voss's MasterClass was the main reason I signed up for MasterClass, and I finished it within a day after signing up. I liked it quite a bit then. Over time, and after having thought about its ideas and watched other related material, I continue to think the ideas are pretty good.

I do not make claims about how much of the material originated from Voss versus was learned by him from others (including the FBI manual). Regardless of Voss's role as innovator versus peddler, the stuff covered is good.

I have already shared several tidbits of insight I learned with friends, but would like to have a public write-up

Of the ideas I learned from the MasterClass, I've already shared several with friends. However, given that the MasterClass is paywalled, I think it would be helpful to have a public write-up of the ideas that I can link to. A public write-up would also benefit people outside my inner circle of friends.

I broadly think the LessWrong community should be more aware of and engage with these bodies of knowledge

Rationality offers a powerful way of viewing the world, but there are also large bodies of knowledge around human interaction developed by others. Fruitful exploration of these bodies of knowledge can help enhance and deepen our rationality.

Worldview

Outward focus

Negotiation tactics are mostly focused outward: they're focused on how to deal with the other side (known as the "counterpart"), not how to deal with yourself.

A large part of the focus of rationality is inward: how to reason better, how to tame and use your emotions, how to achieve goals. Similarly, many techniques such as meditation and relaxation techniques have an inward focus too.

With negotiation tactics, one's own goals obviously matter, but the goals of the other person are front-and-center.

There are a couple of ways that negotiation tactics relate to inward-focused activities. First, in principle it is possible to apply some negotiation tactics in self-negotiation.

Interestingly, some such applications have crude similarities to self-talk techniques (such as CBT or EFT).

Second, in order to negotiate effectively with others, it generally helps to have taken care of yourself *before* that. The idea is that in a negotiation, the counterpart and the situation take center stage. If you're bringing your own baggage to the situation, though, it becomes harder to apply best principles of keeping the focus on the situation and making the counterpart feel heard and addressed.

The negotiation MasterClass does not discuss the first point; it does bring up some ideas related to the second point but indirectly.

Dealing with people who do not aspire to epistemic virtues

The negotiation techniques in the MasterClass grew out of efforts to deal with hostage kidnappers, and evolved to address difficult business negotiations. These people aren't necessarily irrational, but they are not aspiring to epistemic virtues. Pointing out their cognitive biases will not make them thank you for helping them achieve their epistemic goals.

The negotiation techniques therefore can be quite different from norms recommended for rationalist cultures, where there is an assumption that all parties are aspiring toward epistemic virtue. For instance, [tell culture](#) can be great as a rationalist norm but a naive application of tell culture principles would contradict many aspects of negotiation techniques (though reconciliation is possible).

One concern we might have about negotiation techniques is that they specifically *rely* on the irrationality of one's counterparts, and so would fall flat or even backfire in more epistemically virtuous environments. After thinking about the various techniques, I think this actually isn't much of a concern, as long as the person applying the negotiation techniques adjusts the mix effectively based on what is needed. Broadly, I do not think these negotiation techniques are what is sometimes called the [dark arts](#) on LessWrong (though they probably could be interpreted as such!). Later in this post, as I discuss each negotiation technique, I discuss what aspects or variants of it I consider "dark arts"y.

Emotion-focused or cognitively focused?

While Voss gives a lot of importance to emotions, I think it would be wrong to think of the negotiation techniques as primarily emotion-focused. Rather, I think the negotiation techniques try to pick at underlying motivations, some of which could be emotions! I think there is a moderately sound cognitive and epistemic grounding for the negotiation techniques, though the role of human emotions is very important to understanding how successful the techniques will be.

Greater applicability to synchronous interactions

Many of the negotiation techniques are applicable to synchronous interactions, including in-person and phone interactions. In general, the techniques seem optimized for high-bandwidth communication with frequent back-and-forth. I found some techniques as well as some principles to have broader applicability, including to more asynchronous and one-to-many communication contexts.

Low-level execution focus rather than domain-specific tactical or business school-style strategy focus

In the comments, Adam Zerner [writes](#):

It seems worth mentioning that leverage is hugely important. Both 1) having it, and 2) understanding it. For example, suppose you are a programmer applying to companies. 1) It's helpful to be good at interviews and have a lot of companies interested in you. 2) It's helpful to be aware of this fact, and to be aware of what sort of leverage the companies have. ie. BATNA.

Maybe you can call what I am referring to as hard skills and what you are referring to as soft skills? I feel like that isn't a great way to categorize is, but nothing better is coming to me. Whatever the categories are, I think it would be good to explicitly mention that this article is targeting one of them, and that there are other things that are important for the bigger picture of being able to negotiate well.

I also got a related [comment](#) on my Facebook share of the post:

Also if I may seek more details about the Masterclass , does it also delve into concepts taught in management schools (albeit with a different name) like ZOPA(Zone of Possible Agreement) & BATNA (Best Alternative to Negotiated Agreement)

These are great points. There is some discussion in Voss's MasterClass about understanding who has leverage. He also goes into a few nuts-and-bolts bargaining tools like the Ackerman method and provides general advice on price negotiation. He also has a mock job negotiation.

Nonetheless, for the most part, the aspect of negotiations that Voss covers has much more to do with the low-level execution of *how* to share and receive information, as opposed to the *what* of communication. And to the extent that Voss covers other angles, my review skips over them.

Obviously, the *what* of negotiation matters a lot, and there's great advice online around it -- this post does not compete with such advice. In fact, when a lot of people think of negotiation, they're focused on the *what* aspect -- what price should I settle for?

There are a few reasons I think Voss doesn't focus on these much in his MasterClass, and my review here doesn't talk about it at all:

- The *what* is highly domain-specific, and even within a domain requires a fair amount of market research and even getting into the nitty-gritties after your counterpart gives you situation-specific information. A 3-hour MasterClass intended to appeal to a wide audience can't really get into the *what* too much. With that said, in the MasterClass and in several of his interviews that I've linked to, Voss talks about some of the nuts and bolts of the *what* question in the context of real estate negotiation, business partnerships, and job interviews. He's often relying on information he learns back from people who come to the Black Swan Group for training and apply the ideas in real life.
- Highly prescriptive approaches to the *what* question are in tension with the thrust of the approach that Voss is trying to push for -- namely to be open and

curious and let the other side reveal more information to enable collaborative problem-solving. In some of his interviews Voss talks about how he thinks some of the techniques like BATNA are not that useful, but he has a lot of respect for Roger Fisher, who championed these techniques. Voss thinks Fisher's success comes not so much from the techniques as from the emotional intelligence he has when applying them. Voss thinks his own techniques come closer to what needs to be done execution-wise to achieve those sorts of results.

- In a video in the MasterClass session, Voss talks a bit more about his main issue with BATNA, which is that there are many situations where your alternatives are actually pretty bad. This probably comes partly from his own background as a hostage negotiator, where the alternative to negotiated agreement can literally be fatal. So, while it's good to have alternatives, a good foundation of negotiation techniques is one that can work even in the situations where you *don't* have alternatives.

Negotiation principles

This section covers my own interpretation of key "principles" behind the individual negotiation techniques. The way I frame it doesn't always match how Voss presents ideas in the MasterClass or elsewhere, and may not be endorsed by Voss.

Show, don't tell, that you are listening and collaborating in good faith

A lot of the negotiation techniques boil down to a [show, don't tell](#) approach of demonstrating good faith. Many of these "show"s are constructive proof that would be hard to fake for somebody who is not listening and not interested in a good faith collaboration (examples include labeling and the accusations audit). Many of them also directly create value by making substantive progress in zeroing in on the issues involved.

A key point: the "show, don't tell" applies at the meta level of your sincerity and competence; obviously there will be cases where you have to tell the other person factual information or ideas. This is only loosely related to pedagogy's "show, don't tell", so actions that are "show"s in terms of demonstrating sincerity could be "show"s or "tell"s at the object level.

My thoughts on demonstrating good faith

I think this is good advice in general. I do think there are cases where "telling" works, but telling is much more likely to backfire than showing. So I'd say one must always show, but whether to tell depends on how much trust has already been built.

It's about your counterpart

An important aspect of the negotiation techniques is to center them on your counterpart, i.e., the person you are negotiating with. Things like, be curious about what they have to say, be tactically empathetic, always show that you care about the impact of your words and actions on them. Let them go first. By and large, avoid first-person pronouns, so don't say "What I'm hearing is ..." or "I want ..." or "I need ..."

(this is not an absolute injunction, but good to start with). Don't make the other person feel you are putting them on the spot.

My thoughts on counterpart focus

While I think this makes general sense, there are important considerations of asymmetry here that deserve a more detailed treatment. I cover the asymmetric nature of counterpart focus in a later section.

Slow people down and trigger deeper, reflective thinking

Often, in high-stress situations, people are thinking impulsively, defensively, and carelessly. The many negotiation techniques Voss teaches are designed partly to get your counterpart to slow down, relax, and think more reflectively (Kahneman's "system two" thinking). This frame of mind is more conducive to solving challenging problems collaboratively. An additional side-effect of these techniques is also to slow down your own thinking and make you a calmer, better thinker.

My thoughts on slowing down

I'm generally in favor of slowing down and thinking deeply and reflectively. This doesn't translate to being slow in absolute terms -- speeds could vary a lot based on context and familiarity. But I do think it's important to avoid "rushing" things and to combat the tendency for stressed, anxious thought patterns.

Start low, end high

A general theme in negotiation is how both parties perceive it, and whether, at the end of it, they feel like it was a worthwhile endeavor. Starting with the difficult portions and gradually making progress to end on a positive note is important. Voss says that "the last impression is a lasting impression" and emphasizes that much of the positive messaging we are inclined to use to open an interaction may be better suited to the close.

My thoughts on start low, end high

I definitely agree with this. One of the things that makes interactions stressful is when people keep dropping bombshells throughout the conversation. This keeps the other side wary throughout the conversation. Getting the tough parts in the open quickly makes the rest of the interaction more relaxed.

I also agree with the importance of ending on a positive, collaborative note.

Focus less on being in control or being in charge, and more on having the upper hand

Voss says that people often fight to be in control or in charge, e.g., to be the ones speaking in the room or having more overt control of the situation. By relinquishing one's own desire for control, and letting the other person take charge -- while collaborating with them in the process, you can acquire the upper hand by getting more information. Several of the techniques discussed later, such as mirroring,

labeling, mislabeling, dynamic silence, and calibrated questions help the other side feel more in control while also giving you the upper hand by learning more.

My thoughts on control versus upper hand

While I do think being in control is overrated, and it's often more important to learn more than to be in charge, some aspects of this framing didn't resonate much with me. The "upper hand" framing is a little bit in tension with the whole idea of negotiation as being helpful to both sides.

As I discuss in a later section, it can be counterproductive if people start competing to *not go first* -- just as it can be counterproductive if people are competing to go first. So my main takeaway from this point is that if you have a tendency to want to go first and dominate a situation, rethink that. But don't be too singularly focused on *not* going first in all situations.

Negotiation techniques

Mirroring

Mirroring is the technique of repeating about 1 to 3 words of the last sentence the other person said.

Mirroring in general is helpful as it is a low-effort way of showing the other side that you're listening and engaged with what they say.

Mirroring with upward inflection (i.e., a kind of questioning tone) is helpful as a prompt to get the other person to continue expanding and elaborating. This can be helpful if you don't quite understand what the other person said, or you want them to elaborate more.

Instead of the MasterClass, you can watch a [free Chris Voss YouTube video](#) on mirroring.

My personal experience with mirroring

I have not tried to use mirroring much in my life. My impression has been that mirroring is most useful as a low-effort way to engage another person and learn more about what's on their mind. It may be a bit less useful in cases where other, more high-effort and high-reward techniques, can be used.

I used mirroring once in a low-stakes situation with success; a colleague and I were on a call with a third party who ended up not showing up, so we were just waiting for about ten minutes. My colleague was just chatting about stuff going on with his work. I had no particular agenda in terms of information I wanted to know, but I also didn't mind hearing him, so I decided to try using mirroring to help show I was engaged without putting in a lot of effort. This seemed to work well; at the end of it, my colleague said it was great chatting, despite me basically saying nothing.

How "dark arts"y is mirroring?

Mirroring as information-gathering doesn't seem dark arts-y at all to me.

Mirroring as a way of showing you're listening can be dark arts-y if you're *not actually listening*. Somebody practiced enough at mirroring could probably do it automatically without paying close attention to what the counterpart is saying. The counterpart then thinks you were engaged and listening (because they hear your mirrors and don't even realize you were mirroring) but you actually weren't. To be clear, this is not the sort of mirroring that Voss encourages; he emphasizes genuine curiosity and interest.

Does mirroring have a place in rational discourse?

I think mirroring has a place in rational discourse, but a relatively small one. I'm much more excited about the other techniques discussed (that are both high-effort ad high-reward), including labeling.

Labeling

Labeling is the act of providing a short summary of the underlying emotions, thoughts, and ideas behind what your counterpart is saying or doing.

A tactical aspect of labeling: Voss recommends using "It seems/sounds/feels/looks like ..." or "You seem/sound/look like ..." before articulating the label (in a single sentence where possible), and being deferential in tone. It is also open to correction (see the next section, mislabeling). The use of first-person pronoun framings such as "I think ..." or "What I'm hearing is ..." is discouraged because you want it to be about your counterpart and the situation, not about yourself.

Labeling is more high-effort than mirroring: rather than just using short-term memory to remember and pick words from the last sentence, you need to listen to the entirety of what the other person is saying (as well as tone of voice and nonverbal cues where applicable) and summarize it. But it's also more high-reward, because it shows a deeper understanding, helps provide clarity to both sides, and can lead to real progress.

There's a [video](#) going over mirroring and labeling.

Labeling negatives

Voss claims that labeling negatives "always" diffuses them. The key point is to label a negative by accepting it, rather than by denying or contesting it.

For instance, instead of saying "I don't want to sound too demanding" you say "This is going to sound demanding". Or instead of saying "Don't get upset" say "It seems like you are upset".

Instead of the MasterClass, you can check out [this video](#) on labeling negatives.

Labeling positives

Voss claims that labeling positives tends to reinforce them (the opposite of the impact of labeling negatives). For instance, "you sound really excited about this" or "it sounds like you're very happy with the way this turned out".

My personal experience with labeling

I've generally found labeling to be more useful than mirroring, particularly when a lot of raw information is being conveyed and it needs to be processed.

How "dark arts"y is labeling?

Labeling generally seems like a positive thing and not like a dark art. While labels could be misused, I don't see the potential for misuse as big enough to make labeling a dark art.

Does labeling have a place in rational discourse?

Absolutely! I think labeling is pretty valuable. A lot of rational discourse already follows similar practice; for instance, providing labels/summaries of what another person said is viewed positively in the community.

The rationalist community also places more importance on self-labeling, something that negotiation strategy frowns upon (it's not about you, it's about your counterpart).

Mislabeling

Mislabeling refers to the (usually intentional, or at least probabilistically intentional) application of an incorrect or exaggerated label to give the counterpart the opportunity to correct you and reveal more information. So instead of directly asking for the reasons for something, a mislabel might attribute a reason that's probably wrong, and let the other person correct it.

For instance, if a person declines to do an activity, you may mislabel it as "it sounds like you really hate doing this" which gives them the opportunity to correct by saying "actually the time doesn't work, I would love to do it next time".

Brandon Voss (Chris Voss's son and business partner) has a video on mislabeling [here](#).

How "dark arts"y is mislabeling?

There is a "dark arts" form of mislabeling, where your goal is not to get information from the other person, but rather to make them claim something that's beneficial to you. That happens when your mislabel is done with a desire to have the other person reassure you. Mislabeling, when done with curiosity, deference, and a genuine openness to being corrected, and with the goal of getting information rather than manipulate the other person into claiming something, seems good to me.

Does mislabeling have a place in rational discourse?

This is a little unclear. I think mislabeling is okay in rational discourse if the mislabel is still the leading individual candidate. For instance, if you have several hypotheses to explain something, and the leading hypothesis has 30% chance, higher than any other, formulating that hypothesis as a label seems reasonable (and it's probably a mislabel because there's a 70% chance it's wrong). On the other hand, deliberately choosing a low-probability hypothesis as a mislabel seems not very rational.

Dynamic silence

Dynamic silence (also called effective pause) is the idea of just staying silent, usually either right after you mirror, label, or mislabel, or when things fall silent in general. The idea is to give the other person space to keep going on and to correct you. According to Voss, dynamic silence works best after you have established, through mirroring and labeling, that you are listening, engaged, and understanding.

There's a video discussing dynamic silence [here](#).

My personal experience with dynamic silence

I have found dynamic silence to be generally useful when the other side is reasonably articulate. I haven't really needed to consciously practice it; it seems to come naturally.

Calibrated questions: using "what" and "how" rather than "why"

Calibrated questions (also called open-ended questions) are what/how questions designed to both elicit information from the counterpart and start the collaborative process with them by introducing to them the considerations and challenges on your side (Voss calls this "forced empathy"). According to Voss, the main purpose of calibrated questions isn't necessarily to get clear answers, but to get the other side to slow down and think. Two of his favorite questions are:

- "How am I supposed to do that?"
- "What's going to happen if I do that?"

There's a video of Chris Voss explaining these questions [here](#).

Voss is generally against the use of "why" questions because he claims that, universally across languages and cultures, it gets people defensive. When people hear "why" it sounds to them like you think there's something wrong, and they want to defend themselves against accusations. That's because about half the time people use "why" they are being accusatory, and the other half, they genuinely want to know -- and a person on the receiving end can have their defensiveness triggered even when it shouldn't because they have uncertainty about the asker's state of mind.

Legitimate questions

Another related idea that Voss goes over in the MasterClass is to ask "legitimate" questions (and/or raise legitimate concerns) -- basically, the calibrated questions that, if the other side were to hear, they would have to agree that this is a valid point. For instance, rather than asking "How will you guarantee me that X?" ask the underlying question which might be "How do we deal with the fact that I am operating with uncertainty regarding X?" Voss gives the example of the proof-of-life question that hostage negotiators must ask the kidnapper; he suggests a "How do I know the victim is alive?" rather than demanding a proof of life.

My personal experience with calibrated questions

I have generally found the calibrated questions idea to work well, though I've generally used it more for information-gathering than for forced empathy. Avoiding

"why" in particular seems to make a lot of intuitive sense and I've generally found it to be effective at reducing both the defensiveness of the other side and the perceived antagonism in the interaction to third parties.

How "dark arts"-y are calibrated questions?

I don't see any clear mechanism by which calibrated questions are dark arts-y.

Do calibrated questions have a place in rational discourse?

Yes! In fact, I think the idea of using what/how questions instead of why questions has a purely rational basis, in addition to the reasons based on the emotional triggers that "why" could produce. Namely, a "what"/"how" question tends to be more constructive or specific. A "why" question is much more vague, like an injunction to "explain yourself".

For instance, if you ask "Why did you do X?" it's a very difficult question to address. Instead, a question like "What was your motivation for X?" has an appropriate level of focus on motivation. On the other hand, a question like "What events led you to do X?" has a focus on history. By picking a specific aspect, it avoids the very vague "explain yourself" character of the "why" question.

Accusations audit

The "accusations audit" is a comprehensive list of the negative assumptions, thoughts, and feelings that the other side has about you -- both things that they already might harbor and things that they are at risk of harboring once you start revealing the information you're planning to reveal. The accusations audit is then communicated to the other side to show that you're aware of the issues and also to get out in front of them so that both sides can be more relaxed and make progress.

The accusations audit is a proactive version of labeling negatives, where instead of waiting for negatives to pop up before labeling them, you get out in front of them. Some of the remarks made previously about Voss's claim that labeling negatives diffuses them also apply here.

As with labeling negatives in general, it's important in the accusations audit to not try to preemptively defend against or contest the negatives. The first step is to acknowledge the negative, and let the other side acknowledge your acknowledgement. In many cases, the act of acknowledgement itself diffuses the negative enough; in other cases, there's enough time after the acknowledgement to address the substantive issues raised by the negative.

Chris Voss explains the accusations audit [here](#).

Brandon Voss has a practical video on the accusations audit [here](#).

My personal experience with the accusations audit

The accusations audit can be tricky without a good mental model of the other person. It can also be tricky to do if you have a lot of self-esteem issues of your own. I think a key ingredient to the success of the accusations audit is the ability to genuinely think from the other side's perspective, rather than project your own insecurities.

With that caveat, I've found the accusations audit useful, particularly for being able to start discussions that may otherwise not have happened (unlike reactive labeling, the accusations audit can be started asynchronously and used as input to try to get the other side engaged).

How "dark arts"y is the accusations audit?

I don't think the accusations audit is "dark arts"y. With that said, a variant of it could be, where you trump up exaggerated accusations with the goal of manipulating the other person into consoling you. The key way to overcome that is that the "accusations" have to be the other person's accusations toward you, not your accusations toward them. There should be no judgment from your end conveyed as you describe the accusations.

Does the accusations audit have a place in rational discourse?

I think so! I think bringing out negative assumptions, thoughts, and feelings early creates the scope to actually address them in the interaction.

No-oriented questions

The idea behind no-oriented questions is to frame yes/no questions in a way that the answer that will make your life easier is the "no" answer. I found this one of the most interesting and thought-provoking negotiation techniques.

For instance, instead of asking "Is it ok to publish this?" ask "Are there any concerns with publishing this?" Or, in a sales call context, instead of asking "Do you have a few minutes?" ask "Is now a bad time?"

The simplest reason, as given by Voss, is that "yes" can make us feel trapped, like we're being pressured or tricked into agreeing with something. On the other hand, "no" makes us feel freer and safer, like we have protected ourselves. I think there are a few other factors that make no-oriented questions valuable, that I explore in the next few paras.

Chris Voss has a video on no-oriented questions [here](#).

My personal experience with no-oriented questions

Among all the negotiation techniques, I feel that the switch to no-oriented questions has been the one that's influenced my day-to-day actions the most. Part of this might be that this technique is highly applicable to low-bandwidth, asynchronous interaction, that forms a large proportion of my interaction.

In particular, the majority of my questions, particularly around getting consent/approval for joint actions, are now no-oriented. While I think in most cases it doesn't end up mattering, I feel like it does make things smoother when it does matter (i.e., when the other side's answer is the opposite of what's easier for me).

Do no-oriented questions have a place in rational discourse?

Other than the psychological benefits of "no" making us feel more protected, are there other benefits to no-oriented questions? I think so, and I think it would be good to generally push more toward no-oriented questions.

First, I think the act of formulating a no-oriented question helps us, as question-creators, think of the nature of the objection/challenge more clearly. The act of formulating the question itself therefore improves the likelihood of catching issues.

Second, formulating a no-oriented question makes the other side more comfortable responding in *either* way: with a yes (i.e., the answer that makes life harder) *and* with a no (i.e., that makes life easier). By using a no-oriented formulation, you're signaling to the other side that you are prepared for bad news (in the form of a "yes"). This also means that if they deliver bad news, it feels more collaborative -- both sides are actively trying to unearth the bad news and address it rather than hide from it -- and less of a fight.

Third, responding to a no-oriented question also triggers more thinking from the other side, partly because you've already shown that you are willing to hear bad news.

Summaries (available only in the session)

The "summary" is a technique not covered explicitly in the regular MasterClass, but covered more in the session. A summary is usually a paragraph or two that summarizes what the counterpart has been saying and/or what all has been discussed so far.

The summary is something that should be used once you've extracted relevant information using mirrors and labels. A summary can be somewhat similar to labels, in that individual sentences in the summary can be similar to labels, but it differs in that it puts everything together.

Summaries could also be used at the beginning to start with what's known already, but it's usually better to begin with an accusations audit if that applies, and to wait till there's more information before attempting a summary.

"You're right" versus "that's right"

Voss says that getting the other side to say "you're right" isn't great, but getting the other side to say "that's right" is great. "That's right" is an acknowledgement by the other side that you've understood the situation.

While "that's right" is great to hear at any time (e.g., in response to a good label), the summary is the time in the conversation when you're most likely to get it, because a good summary that exactly speaks to the other side gets the most satisfied response from them.

My personal experience with summaries

I had not come across summaries as a negotiation technique until I watched the session in December 2021. I have done summaries in the past, and generally think they are useful.

I have a little caveat / nuance to add to the "you're right" versus "that's right" distinction. While "that's right" is almost always a good sign, "you're right" isn't always a bad sign; the tone of voice and the context matters. If somebody says "you're right" right after you go on a tirade, it's probably a bad sign and suggests they may want you to shut up. However, if the counterpart has a spontaneous epiphany and says "you're right about X!" that's good. So the context and tone of voice matter a lot in interpreting "you're right."

Paraphrasing (available only in the session)

Paraphrasing, like mirroring, involves repeating the things your counterpart said back to them. However, paraphrasing isn't just the last few words, but goes over a larger portion of what they said, picking out key ideas. Paraphrasing doesn't need to use the exact words, but it should still retain the meanings.

Paraphrasing is similar to summaries but involves less *synthesis* and more just a literal reading and compression of what was said.

When your counterpart is unloading a lot of information, interrupting them to paraphrase it, and then handing the control back to them, might be a good idea. While people don't like being interrupted, so they may feel a little bit of irritation in the moment, their fear that you interrupted them just to say your own thing will dissipate once you hand control back to them.

Roughly speaking:

Mirrors : Paraphrasing :: Labels : Summaries

My personal experience with paraphrasing

I have not used paraphrasing much, but generally agree with some of the benefits.

Using "that's my problem" when introducing constraints (only in the session)

This is a little related to the accusations audit, but a slightly different tack. Sometimes, in a negotiation, you need to introduce a tricky constraint from your side. If you expect that the first reaction of the other side to you introducing the constraint would be for them to think "that's your problem, not mine" then -- preface with "that's my problem" before introducing the constraint. That way, it becomes clear to the counterpart that while you are seeking their help to work within the constraint, you aren't allocating blame or responsibility for the constraint to your counterpart.

My personal experience with "that's my problem"

I like this idea. I hadn't specifically encountered it before the MasterClass session, and haven't had many opportunities to apply it, but I expect I will.

Other ideas

Tactical empathy

This is not so much a single technique as an underlying idea that influences other techniques; nonetheless it's close enough to a technique. The idea behind "tactical empathy" is to demonstrate -- through actions -- that you understand and respect the situation your counterpart is facing and the impact it's having on them (of course, if you don't yet understand, things like mirroring, labeling, and calibrated questions help get you there). Tactical empathy is different from but related to sympathy ("I feel how you feel"). The goal of tactical empathy is to get the other side to trust that you have a good enough understanding of their situation that you can collaborate with them to solve problems.

There is a lengthy video by Voss on tactical empathy [here](#).

Trust-based influence

In the MasterClass and elsewhere, Voss uses the term "trust-based influence" to describe the kind of influence that you can build through the use of tactical empathy, supported by other methods discussed earlier (mirroring, labeling, calibrated questions). With trust-based influence, the other side understands that you understand their situation and respect them, and therefore trusts that you'll be able to collaborate with them to solve problems.

My personal experience with tactical empathy

I have not actively applied tactical empathy as a technique, but I think my practice has moved a bit in that direction after being exposed to the concept.

How "dark arts"y is tactical empathy?

I think tactical empathy can be "dark arts"y if you actually don't understand the counterpart's situation.

It's still likely less "dark arts"y than false sympathy or false agreement.

Does tactical empathy have a place in rational discourse?

Tactical empathy seems closely related to proto-rationalist ideas such as [Rogerian argument](#) and the similar "[ideological Turing test](#)".

Tone of voice

Voss identifies three kinds of tones of voice:

- Assertive: Ideally never use this!
- Playful/accommodating: Use this when learning and collaborating. Generally, this should be used about 80% of the time.
- Late-night FM DJ: This is a calm, slow, and firm tone used to communicate immovability. Use this instead of the assertive voice when standing firm. Generally, this should be used about 20% of the time.

He also identifies two kinds of inflections:

- Inquisitive (upward inflection)
- Declarative (downward inflection)

Inquisitive inflections are good when you want to get the other side to talk more.

My personal experience with tone of voice

Since a lot of my communication is text-based rather than voice-based, these concepts have had limited utility to me. One general idea that I've taken from this is to speak more slowly (one of the aspects of the late-night FM DJ voice). I might also have reduced my use of assertive voice and increased my use of playful/accommodating voice as a result of being influenced by these ideas.

Some of the principles also carry over from tone of voice to tone of text communication. Even prior to this MasterClass, I adopted a upbeat style of communication, sprinkling exclamation points and smileys in future drafts. I've continued with this practice.

How "dark arts"y is tone of voice?

I don't see anything "dark arts"y about using a playful/accommodating tone of voice, or the late night FM DJ voice.

Does tone of voice have a place in rational discourse?

Tone of voice (or tone of text) provides another dimension of communication that influences discourse. Using it in a good way seems consistent with the idea of rational discourse. Since a lot of rational discourse is centered around open exchange of ideas, the playful/accommodating tone of voice seems suited to it.

Tackling loss aversion

Voss talks about [loss aversion](#) -- he calls it "fear of loss" -- and says that this distorts people's thinking a lot. So what do you do if the idea on the table isn't about a loss? His suggestion seems to boil down to framing it in terms of a loss by using the [opportunity cost](#) framing: basically point to the other side what they lose by *not* doing the deal.

Voss talks about loss aversion in [this video](#) where he cites the academic field of [prospect theory](#). He says that reframing a forgone gain as a loss can be so powerful that the term *bending reality* can be used for it.

How "dark arts"y is tackling loss aversion?

To the extent that this is about combating and neutralizing an existing loss aversion bias, I think it's not a dark art. But to the extent it's about invoking loss aversion and creating distortionary fear, I think it is a dark art. A lot depends on the implementation.

My personal experience with tackling loss aversion

I do not remember any situations where I consciously applied this technique; also, I was already broadly aware of the ideas of loss aversion and opportunity cost so the marginal impact of the MasterClass was low for me.

Black swans

The term "black swan" refers to a hidden piece of information that, once revealed, changes the shape of the negotiation. Part of the goal of negotiation is the (collaborative) discovery of these black swans. This can be thought of as one reason it's so important to get the other side to talk and reveal more private information (combining this private information with your private information can help unearth the black swans).

My personal experience with black swans

I have not had any major success unearthing black swans! But I still think it's a valuable idea.

How "dark arts"y are black swans?

I don't see the idea of trying to discover black swans as "dark arts"y.

Do black swans have a place in rational discourse?

Absolutely! I think a lot of rational discourse is about discovering new ideas, and some of the more novel ones could qualify as black swans.

Fairness and reciprocity

The related ideas of fairness and reciprocity come up a lot in negotiation. Awareness of these can help.

Voss talks of a few ideas related to these:

- Avoid triggering reciprocity by e.g., making asks/demands, when there are other alternative ways: For instance, the use of legitimate calibrated questions can engage your counterpart to collaboratively solve the problem with you, without triggering the sense that they are doing you a favor.
- Rather than say things like "I only want what's fair" (that can be read as an accusation of unfairness) reassure the other side that if at any point they feel that they aren't being treated fairly, they should speak up.
- Offer things to the other side (this could include goodies or information) that aren't costly for you, but that are either directly valuable to the other side, or at least signal that you are there to help them and/or that they are squeezing out good value from you. An example is mentioned in the final stage of Ackerman bargaining: once you are at the limit of the budget you are willing to pay, offer some non-monetary good that is cheap for you -- and may even be something not valuable to the other side (if you can't think of anything valuable to them) -- but that shows the other side that you are stretched to the limit with the money side and they've gotten a good deal.

Negotiating styles (covered more in the session)

In the MasterClass session, Voss and colleagues talk about three negotiating styles/personalities; you can see more details [here](#):

- Assertive style/personality: The assertive style is to assert, to confront, and to make sure one's point of view is heard.
- Analyst style/personality: Analysts care more about getting data and information, and tend to be reticent to say much, especially early on.
- Accommodator style/personality: Accommodators want and try to be liked; they value harmony and feel the need to diffuse any tension or awkwardness.

The personalities correspond to the tones of voice previously discussed; assertive style corresponds to assertive tone of voice, analyst style corresponds to the "late night FM DJ" voice, and accommodator style corresponds to the playful/accommodating tone of voice. However, the underlying style and tone of voice don't need to match. For instance, the assertive *tone of voice* should probably never be used, but the assertive style/personality can still be expressed through other tones of voice and can be important.

The general advice for different personalities is to focus on incorporating more of the techniques that don't naturally jive with that personality. Here are some of the ideas if you're one of these personalities:

- Assertive style/personality: Assertives need to make sure they are not using an assertive *tone of voice*. Also, they may need to be more conscious of shutting up and *listening* so as to hear what the other side has to say.
- Analyst style/personality: Analysts need to make sure that they are sufficiently warm and friendly, as the analyst personality can appear too reticent at times. Active application of negotiation techniques such as mirroring and labeling can help with this.
- Accommodator style/personality: Accommodators tend to generally be good listeners, but tend to have more trouble with using tools like dynamic silence to get the other side to talk. They may also end up talking or revealing too much, or overpromising.

There is also general advice for dealing with the different sorts of personalities in your counterpart:

- Assertive style/personality: When dealing with assertives, it's particularly important to let them have their say first. This is important even in general, but less so for analysts and accommodators, who don't generally have that much of an urge to have their say first.
- Analyst style/personality: When dealing with analysts, it's helpful to draw them out more. Instead of asking questions, using labels and mislabels can help draw out the analyst due to their desire for correctness.
- Accommodator style/personality: When dealing with accommodators, it's important to get them to feel more comfortable to voice their own concerns and get out of the mode of trying to accommodate. Also, accommodators may tend

to overpromise when they feel pressure to please, so it's important to be careful about that.

Zigzag path to progress (described more in the session)

One general idea presented in the session, and demonstrated in various demonstrations, is that the path to progress isn't always monotonic -- there can be some ups and downs along the way. One of the mechanisms is that as people get more comfortable in the negotiation, they reveal more of their challenges and negative feelings, which can sometimes cause them to heat up. Also, sometimes the application of labels and mirrors may be off; for instance, a mislabel can cause the counterpart to get more angry sometimes.

However, even if the emotional journey isn't always one of progress, the informational journey usually is -- even with the rocky emotional moments, new information is emerging that can ultimately be used to make more progress.

General concerns

This section goes into detail on general concerns that I've had or that have been raised to me about negotiation techniques. This includes concerns raised from rationalist perspectives.

Asymmetry between you and the counterpart

While there is a lot of homage paid to the idea of negotiation not being a zero-sum game, the fact is that most of the negotiation techniques are applied asymmetrically between you and your counterpart. For instance, "let them go first" -- what happens if both you and your counterpart are trained in negotiation techniques and trying them on one another?

Another way of thinking of it game-theoretically: are negotiation techniques like defecting in a prisoner's dilemma, achieving gains at the other person's expense, but if both people did it, both sides would lose out?

In one of his interviews, Voss addresses this in some depth. He says that he actually doesn't mind, and even prefers, if people use his negotiation techniques on him. He says, for instance, that most of his co-workers ask no-oriented questions all the time. He's used to it and doesn't consider it manipulative.

What if both sides were using negotiation techniques on each other? In particular, what if both sides are trying to let the other side go first? Could this result in an impasse? Some more advanced treatments of negotiation techniques (including some parts of the MasterClass) discuss this. Generally, if the other side is really passionate about wanting you to go first, it's a good opportunity to learn what concerns they have about going first, for which you can use mirroring, labeling, and calibrated questions. Ultimately, if there's something that's a deal breaker for them, it's good to know that in a way that doesn't involve confrontation and anger, and that's what the negotiation techniques are for.

So, whereas the fundamental criticism that negotiation techniques are asymmetric is true, they can be adapted easily to a symmetric world, and that symmetric world is

likely even better for both parties than if only one side is applying negotiation techniques.

How should awareness of the negotiation techniques affect your expectations of how others interact with you?

The bad, entitled way to use your knowledge of negotiation techniques is to start expecting that people around you will start using the nice parts of them on you. For instance, maybe you realize how great it is to hear calibrated questions instead of "why" questions, or how much better no-oriented questions sound (on the receiving end) than yes-oriented questions. An entitled application of this enhanced knowledge would be to start suggesting/demanding that the people around you start using these techniques with you so that your lived experience is nicer. However, you rarely have the level of control over other people to do this in a big enough way, and it's entitled to expect that they do. I do think it's worth sharing these ideas with friends so that they can be more effective, just not primarily for the purpose of them providing a better experience to you!

A better way to apply your knowledge of these is to come to situations with more awareness of your own subconscious triggers. When somebody asks a why question, and you feel defensive or irritated, notice that, and think about how much of this is the "why" framing of the question. In some cases, the other person's use of "why" might reflect genuine irritation and hostility on their part. In other cases, though, it may be an innocuous word choice. One thing I have found useful is to notice my slight defensiveness at being asked why questions, then pause, and then answer them instead as if I had been asked a corresponding what/how question. In almost all cases, this works really well. In rare cases where I detect extreme hostility in the why question, or extreme lack of clarity in it, some other methods such as mirroring, labeling, or asking calibrated questions back can help.

The same goes with the use of yes-oriented questions: I now tend to notice my sense of feeling pressured when somebody asks me a yes-oriented question that I do not fully want to say yes to. In such cases, I sidestep the use of a binary response and answer in the same sort of way as if I'd been asked a no-oriented question.

In addition to helping me respond better to cases where others ask questions or make remarks to which my initial response is negative, this awareness also gives me a better lens when viewing interactions (written or oral) as a third party. When an interaction that starts off cordially becomes openly antagonistic, or when an interaction seems to have undertones of hostility, I can often locate things like why questions and yes-oriented questions in there.

Insufficient self-expression

The outward focus of negotiation techniques counters a lot of advice on the importance of expressing yourself. If everybody were busy negotiating, would people avoid expressing themselves?

I think this is an important criticism and it's important to remember that in a context where the other side is interested in what you have to say, expressing yourself is good (because it's actually meeting a need of both sides, i.e., it is a [coincidence of wants](#)). Negotiation techniques are more the scaffolding you put around self-expression, they

aren't about the self-expression itself but they help support it by helping create a safe environment for that expression.

For instance, the idea of letting the other side go first and extracting information from them is important because, generally, only once people have had their say do they feel relaxed enough to hear you.

Similarly, the accusations audit helps diffuse the other person's negative valence around you, putting them in a position to actually listen to you.

Are negotiation techniques symmetric weapons, asymmetric in a good way, or asymmetric in a bad way?

Scott Alexander introduced the concept of asymmetric weapons in two blog posts ([here](#) and [here](#)). The first blog post highlighted reason as an asymmetric weapon that generally helps push toward epistemic progress. The second blog post pointed out several ways that the asymmetric weapon of reason could actually make things worse. A question about negotiation principles and techniques is whether they are symmetric, or asymmetric in a predictable way.

I think negotiation principles and techniques are *slightly* asymmetric in a positive direction. Overall, if more people adopted these successfully, interactions would be calmer and more collaborative, and this would lead to better outcomes for all.

Principal-agent concerns

In many cases, your counterpart is an *agent* of an organization or another entity (the *principal*). For instance, you might be negotiating with an employee of a company you are doing business with. The employee is the agent. The business is the principal. In any situation where the agent and principal differ, there is potential for a [principal-agent problem](#).

The bulk of negotiation techniques are focused on the agent, and as such, they may exploit the principal-agent problem to get good deals for yourself. Many of the examples provided by Chris Voss, including free hotel room upgrades, free flight ticket upgrades, and special coupons at stores, seem to be open to this criticism.

I do think this is a valid but ultimately minor criticism. In most of these cases, the principals have already made a macro decision as to how much discretion to grant agents, and your actions are operating within that discretionary framework. For instance, if the hotel front desk staff has the ability to offer you a late checkout, that's because the hotel management decided to grant them this flexibility.

That said, there could be exceptions, and I think as an individual you can decide/choose not to apply negotiation techniques to situations where you feel it's exploiting a principal-agent problem in a particularly bad way.

How much does this move the needle in the real world?

The ultimate criticism is that this all sounds cool, but how consequential is it to the challenges facing civilization? Surely we can't negotiate our way out of AI risk!

I don't have a solid answer, but here is a tentative reason to think this is important. First, improving the quality of cooperation in general -- both in terms of the objective results produced and the positive vibes around it that make it more sustainable -- seems extremely important for tackling difficult challenges. People like Brian Tomasik have [written](#) about the broad theme.

Negotiation techniques done right seem like a good way to improve collaboration, coordination, and cooperation, at least at the micro-level. The handwavy part is getting from that to macro-level improvements in cooperation, in ways that meaningfully improve the world. I don't have a lot of confidence in how strong that connection is. But I think it might be enough to at least give negotiation techniques a try. The magnitude is uncertain but I think the effect is likely positive.

Conclusion

Overall I am glad to have been exposed to the negotiation techniques and ideas popularized by Chris Voss. I think many of them could be valuable to readers on LessWrong. Thank you for reading all the way till here, and please don't hesitate to share your thoughts in the comments here!

Chu are you?

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.
This is a linkpost for <http://adelelopez.com/chu-are-you>

[Maybe you've heard about something called a Chu space around here](#). But what the heck is a Chu space? And whatever it is, does it *really* belong with all the rich mathematical structures we know and love?

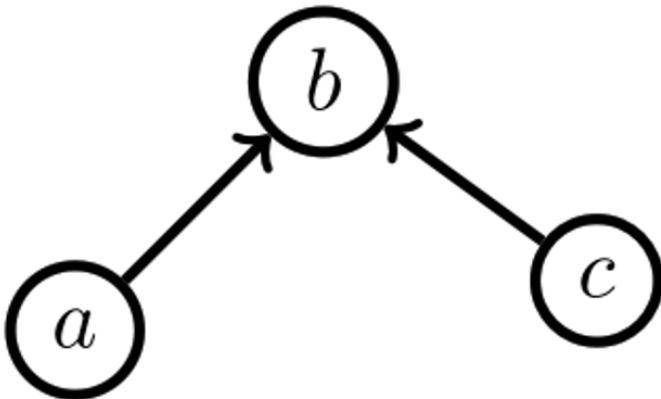


Say you have some stuff. What can you do with it?

Maybe it's made of little pieces, and you can do a different thing with each little piece.

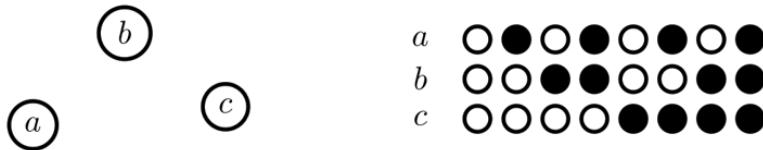


But maybe the pieces are structured in a certain way, and you aren't allowed to do anything that would break this structure.



A Chu space is a versatile way of formalizing anything like that!

To represent something in a Chu space, we'll put the names of our pieces on the left. How about the rules? For a Chu space, the rules are about allowed ways to color our pieces. To represent these rules, we can simply make columns showing all the allowed ways we can color our pieces (just get rid of any columns that break the rules). Here's what a basic 3-element set (pictured on the left) looks like as a Chu space:



It doesn't have any sort of structure, so we show that by allowing all the possible colorings (with two colors). Chu spaces that don't have any rules (i.e. all colorings are allowed) are equivalent to sets.

What about the one with the arrows from above? How can we make an arrow into a coloring rule? One way we could do it is by stipulating that if there's an arrow $x \rightarrow y$, we'll make a rule that if x is colored black, then y has to be colored black too, where x and y can stand in for any of the pieces. Here's what that Chu space looks like:



Spend a minute looking at the picture until you're convinced that our coloring rule is obeyed for every arrow on the left side of the picture. Any Chu space that has this kind of arrow rule has the structure of a poset.

There and back again

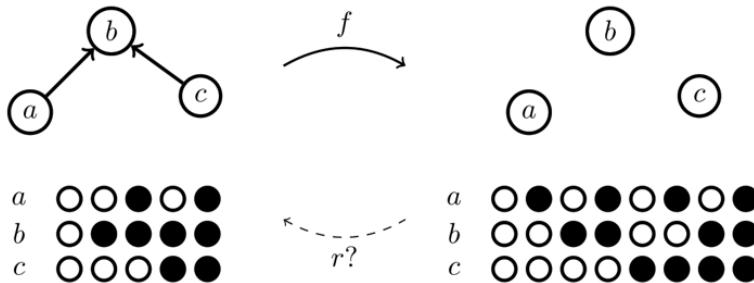
If we have two Chu spaces, say A and B, what sort of maps should we be able to make between them?

We'd like to be able to map the pieces of A to the pieces of B. So this part of our map will just be a normal function between sets:

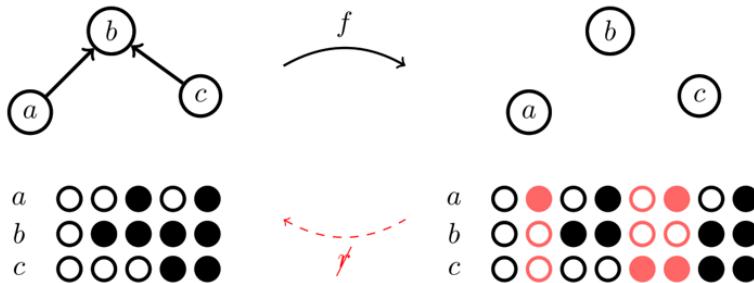
$$f : A_{\text{pieces}} \rightarrow B_{\text{pieces}}$$

But we also want our maps to respect the rules of A as it maps to B. How can we check this?

Let's think through how we would be able to tell if a potential map did break the rules. The map will take pieces of A to pieces of B, so let's look at the pieces of B that got mapped onto, i.e. $f(A_{\text{pieces}})$. It will help if we have a concrete example in mind, so let's consider a potential map that we think should break the rules: one that breaks the arrow structure of a poset by breaking it up into a mere set. For simplicity, we'll just have the pieces map f take pieces to pieces with the same label.



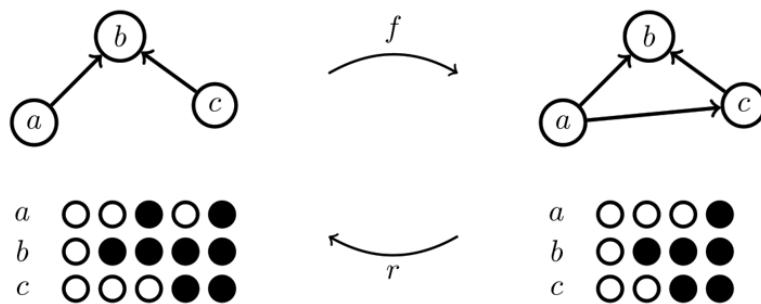
We can see now that if this potential map breaks the rules, it must be because one of the colorings for $f(A_{\text{pieces}})$ is invalid. Specifically, the colorings highlighted in red:



The problem with these colorings is that there are no colorings in the original space for them to correspond to; so they must break one of the rules of the original space. So for a map that does follow the rules, we'll want to make sure every coloring in B has a corresponding coloring in A. This will be another function between sets, this time going backwards between the sets of colorings:

$$r : B_{\text{colorings}} \rightarrow A_{\text{colorings}}$$

Now let's look at an example of what we expect should be a legit map between Chu spaces.



Again, we are just mapping pieces to pieces with the same labels. This map is ok because even though B has some additional structure, it respects all of the structure from A .

The last part we need to define a map between Chu spaces is to determine exactly what it means for colorings of B to have a corresponding coloring in A. For a given piece x , and a given coloring s , we have a function that gives us the color of x using the coloring s :

$$\text{Color}_A : A_{\text{Pieces}} \times A_{\text{Colorings}} \rightarrow \text{Palette}$$

We want to make sure that if we have a coloring s from B, that it gets taken back to a compatible coloring in A. The coloring it gets taken back to is $r(s)$, and we can check its color on any piece p from A with $\text{Color}_A(p, r(s))$.

What do we want this to be the same as? It should be the same as our coloring s from B on all the pieces that f maps onto! We can check these colors with $\text{Color}_B(f(p), s)$. And so, we can finish our notion of a Chu map by making sure that for all the pieces p of A, and for all colorings s of B, that the following equation holds:

$$\text{Color}_B(f(p), s) = \text{Color}_A(p, r(s))$$

Thus, a map between Chu spaces is made of any two functions f and r which satisfy the above equation.

Basic concepts

In order to talk about Chu spaces more easily, let's define some terminology.

The set of "pieces" is known as the **carrier**, and each "piece" is called a **point**. These points index the **rows**. Likewise, each "coloring" (i.e. **column**) is indexed by a **state**, and the set of states is called the **cocarrier**. Each point-state pair has a "color" which is a value taken from the **alphabet** (or "palette"). For a Chu space C with point p and state s , we'll denote this value by $C(p, s)$.

A **Chu transform** is a map between two Chu spaces: $t : A \rightarrow B$. It is composed of two functions, the **forward** function f from the carrier of A to the carrier of B, and the **reverse** function r from the cocarrier B to the cocarrier of A. These must satisfy the **Chu condition** in order for this t to be a Chu transform: For every point p_A of A and every state s_B of B,

$$B(f(p_A), s_B) = A(p_A, r(s_B))$$

We say a Chu space is **separable** if all the rows are unique. It's **extensional** if all the columns are unique, and **biextensional** if both the rows and the columns are unique. We can make any Chu space biextensional simply by striking out any duplicate rows and columns. This is known as the **biextensional collapse**. Any two Chu spaces with the same biextensional collapse are **biextensionally equivalent**. It's very common in applications to only consider things up to biextensional equivalence.

Chu spaces with Chu transforms form a category. This category is typically called $\text{Chu}_{|\Sigma|}$, where Σ is the alphabet.

Representation

Lots of categories are fully embedded into Chu_2 , and even more if we allow arbitrarily many colors. Let's look at some examples so we can get a better feel for how we can represent different kinds of rules with coloring rules.

For a topological space, the pieces will be all the points. The allowed colorings will be exactly the ones where the pieces colored white make up an open set, and the pieces colored black make up a closed set. It's then easy to see how the Chu transform gives us exactly continuous functions between topological spaces!

This example also motivates why Chu spaces are called spaces: for any two color Chu space, we can think of each column representing a generalized open set, containing the white colored points.

If we use more colors, we can embed any category of algebraic objects fully and faithfully into Chu! In other words, Chu can represent any algebraic category. To see how this works, let's see how we can represent any group. The points will be the elements of the group. We'll need 8 colors, which we'll think of as being the 8 combinations of red, green, and blue. We'll think of each relation $rg = b$ as having the r slot colored red, the g slot colored green, and the b slot colored blue. Only colorings which have at least one element in the right color slot for ALL the relations of the group are allowed. (See Note 1 for more details.) We need 8 colors to represent the possibility of the same element appearing in multiple slots. For example, with the zero group, 0 appears in every slot for every relation, so the 0 row will have all the colors which have at least red, green, or blue "turned on":

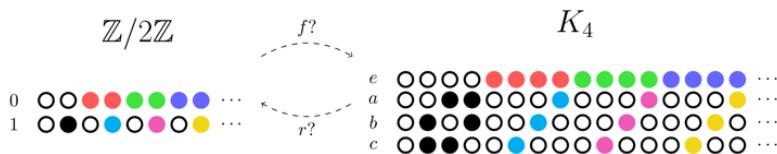
$$0 \quad \boxed{0} + \boxed{0} = \boxed{0}$$



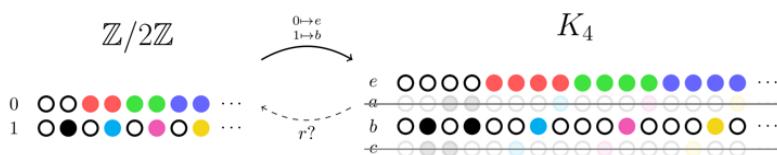
This is not an economical representation. Even for just the cyclic group of order 2, we need lots of rules (41, to be exact).

The diagram illustrates the addition of two vectors from $\mathbb{Z}/2\mathbb{Z}$. The first vector is labeled 0 and consists of 15 zeros. The second vector is labeled 1 and consists of 15 ones. The sum of these two vectors is also labeled 1 and consists of 15 ones. This visualizes how the additive identity (0) and the additive inverse (1) work in the finite field $\mathbb{Z}/2\mathbb{Z}$.

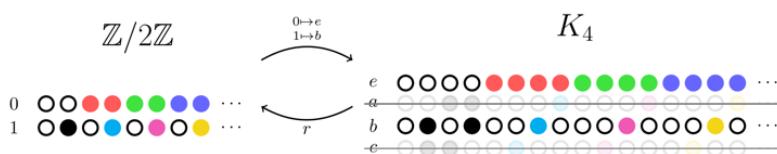
The Klein 4-group requires 1045 columns under this representation! So the following pictures will be truncated, so that we can see the essential idea without being overwhelmed. Let's consider a potential Chu transform between $Z/2Z$ and K_4 with this representation:



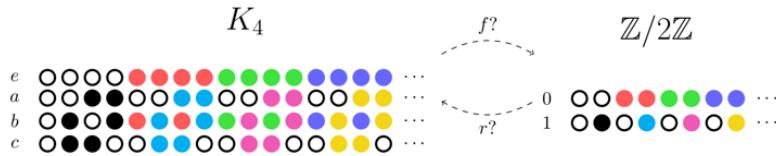
The forward map f is almost like a question. It *chooses* which rows 0 and 1 should correspond to (say e and b respectively), and asks if this is an allowed transformation.



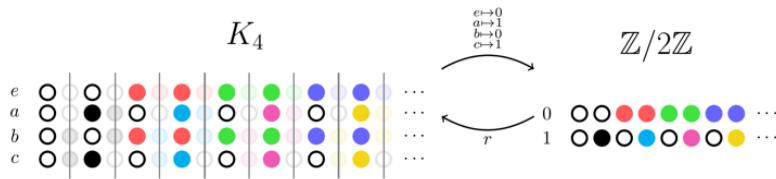
The reverse map r responds by finding the representative of each column from K_4 . E.g. the 2nd column of K_4 must be mapped to the 2nd column of $Z/2Z$. If it can't find such a map, we can think of this as an answer in the negative.



Let's look at another example where the reverse map is slightly more interesting. We expect there to be a group homomorphism from K_4 to $\mathbb{Z}/2\mathbb{Z}$, so let's check that this is the case for the Chu spaces as well. (We'll show a different subset of the columns of K_4 in this example.)



Again, we'll choose a forward map, this time taking e and b to 0, and a and c to 1. The reverse map then verifies that the group structure is satisfied, by looking for a column of K_4 for every column of $\mathbb{Z}/2\mathbb{Z}$.



Notice how the alternating color columns get chosen by r . This is mandated by the Chu condition, given that f maps its rows in an alternating pattern.

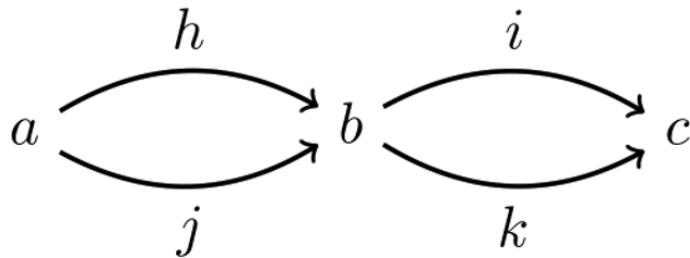
Notice also how we didn't need to add anything else to properly represent group axioms, such as the fact that every group has an identity. Instead, this is encoded implicitly: the identity row will be the only row that doesn't contain any black, so the Chu condition thus ensures it must always be mapped to the identity of another group represented this way. By simply specifying all the relations that are allowed, we've implicitly specified the entire structure! This seemingly innocuous observation is at the heart of the celebrated Yoneda lemma.

Yoneda

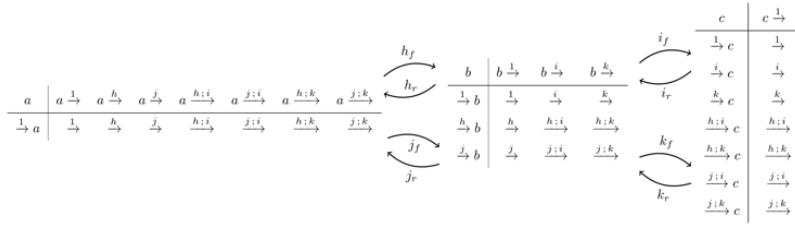
The [Yoneda lemma](#) could rightly be called "the fundamental theorem of category theory". While fundamentally a simple result, it is notoriously difficult to understand. An important corollary is the Yoneda embedding (and also the co-Yoneda embedding). There is an analog of the Yoneda embedding for Chu spaces which can be proved directly (without the Yoneda lemma) and which is conceptually simpler! I'll prove that version in full here, since I found it enlightening to see exactly how it works.

Theorem: Every small (see Note 2) category C embeds fully into $\text{Chu}_{|C|}$.

Proof: The embedding works by defining a functor $\wp : C \rightarrow \text{Chu}_{|C|}$ (\wp is the [hiragana kana](#) for "yo"). For each object $c \in C$, \wp takes this object to the Chu space where the points are all the arrows *into* c , and the states are all the arrows *out of* c . Here's an example category E :



And here are the three Chu spaces that \wp makes out of E (one for each object), along with some of the induced Chu transforms (which will be defined below):



In any case, $\wp(c)$ is separable because the identity column will be different for each distinct point (i.e. incoming morphism).

Similarly, it is extensional since the identity row will be different for each distinct state (i.e. outgoing morphism). So $\wp(C)$ is bixextensional.

Now consider a morphism $g : c \rightarrow d$ in C . $\wp(g)$ must of course be a Chu transform, hence is composed of a pair of functions (g_f, g_r) between the carriers and cocarriers of $\wp(c)$ and $\wp(d)$. Specifically, g_f will take a point p of $\wp(c)$ (i.e. incoming morphism to c) to a point of $\wp(d)$ (an incoming morphism to d) defined by $g_f(p) = p ; g$. Similarly, g_r will take a state s of $\wp(d)$ to a state of $\wp(c)$ via $g_r(s) = g ; s$. This satisfies the Chu condition since function composition is associative:

$$\wp(d)(g_f(p), s) = g_f(p) ; s = (p ; g) ; s = p ; (g ; s) = p ; g_r(s) = \wp(c)$$

This functor is faithful, which means that it keeps morphisms distinct: if $g, h : c \rightarrow d$ are distinct morphisms, then $\wp(g)$ and $\wp(h)$ are also distinct. We can see this by checking the values of $g_f(1_c) = 1_c ; g = g$ and $h_f(1_c) = h$.

And finally, this functor is full, which means that every Chu transform between the Chu spaces of $\wp(C)$ comes from a morphism of C . We can see this by taking an arbitrary Chu transform $(f, r) : \wp(c) \rightarrow \wp(d)$. From the Chu condition $\wp(d)(f(1_c), 1_d) = \wp(c)(1_c, r(1_d))$, which implies that $f(1_c) = r(1_d)$. By construction, this is a morphism m of C starting at c and ending at d , i.e $m : c \rightarrow d$. Again by the Chu condition, $f(p) = \wp(d)(f(p), 1_d) = \wp(c)(p, r(1_d)) = p ; m$. Similarly, $m ; s = (f(1_c), s) = (1_c, r(s)) = r(s)$. Thus, f is exactly m_f , and r is exactly m_r as given by $\wp(m)$. This means our transform was just $\wp(m)$. QED

Having a full and faithful embedding into $\text{Chu}_{|C|}$ means that our category C is *represented* by $\text{Chu}_{|C|}$. This is quite similar to how groups can be represented by vector spaces!

Also notice how we needed three things to make this work: identity morphisms for each object, composition of morphisms, and associativity of composition. These are exactly the requirements for a category! I think this explains why categories are such a fruitful abstraction: it has exactly what we need to make the Yoneda embedding work, and no more.

Final thoughts

Hopefully I've convinced you that Chu spaces are indeed a mathematical abstraction worth knowing. I appreciate in particular how they provide such a concrete way of understanding otherwise slippery things.

This post just scratches the surface of what you can do with Chu spaces. Now that I've laid out the basics, I plan to continue with another post about how Chu spaces relate to linear logic, cartesian frames, Fourier transforms, and more!

Most of the stuff in this post can be found in [Pratt's Coimbra paper](#). There aren't many distinct introductions to Chu spaces so I thought it was worth retreading this ground from a somewhat different perspective.

Special thanks to Evan Hubinger for encouraging me to write this up, and to Nisan Stiennon for his helpful feedback!

Footnotes

Note 1: For $Z/2Z$, there are four relations of the form $r + g = b$. Let's start with $0 + 0 = 0$. To cover this relation, we must turn on either the red, green, or blue light for 0. So the 0 row will never be colored black. If we color 0 white, then we've covered every relation, since every relation has 0 in either the r , g , or b slot. On the other hand, if we colored 0 green, then we would need to turn on either the blue or the green light for 1 in order to cover the second equation $0 + 1 = 1$. If we turned on the blue light for 1,

then the third equation $1 + 0 = 1$ would be covered already, but the fourth equation $1 + 1 = 0$ won't. We'll have to turn on either the red or the green light for 1 in order to cover the fourth equation. [Here's the code](#) I used to calculate these. ↩

Note 2: A *small* category is simply one where the objects and morphisms are both sets. They could even be uncountable sets! A category that *isn't* small is Set. That's because the objects of this category are all sets, and we can't have the set of all sets lest we run into [Russell's paradox](#). ↩

AI Tracker: monitoring current and near-future risks from superscale models

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

This is a linkpost for <https://aitracker.org/>

TLDR: We've put together a website to track recent releases of superscale models, and comment on the immediate and near-term safety risks they may pose. The website is little more than a view of an Airtable spreadsheet at the moment, but we'd greatly appreciate any feedback you might have on the content. Check it out at aitracker.org.

Longer version:

In the past few months, [several successful replications](#) of GPT-3 have been publicly announced. We've also seen the [first serious attempts](#) at scaling significantly beyond it, along with indications that large investments are being made in commercial infrastructure that's intended to [simplify training](#) the next generation of such models.

Today's race to scale is qualitatively different from previous AI eras in a couple of major ways. First, it's driven by an unprecedentedly tight feedback loop between incremental investment in AI infrastructure, and expected profitability [1]. Second, it's inflected by nationalism: there have been [public statements](#) to the effect that a given model will help the developer's home nation maintain its "AI sovereignty" — a concept that would have been alien just a few short years ago.

The replication and proliferation of these models likely poses major risks. These risks are uniquely hard to forecast, not only because many capabilities of current models are novel and might be used to do damage in imaginative ways, but also because the capabilities of future models can't be reliably predicted [2].

AI Tracker

The first step to assessing and addressing these risks is to get visibility into the trends they arise from. In an effort to do that, we've created **AI Tracker: a website to catalog recent releases of superscale AI models, and other models that may have implications for public safety around the world.**

You can visit AI Tracker at aitracker.org.

Each model in AI Tracker is labeled with several key features: its input and output modalities; its parameter count and total compute cost (where available); its training dataset; its known current and extrapolated future capabilities; and a brief description and industry context, among others. The idea behind the tracker is to highlight these models in the context of the plausible public safety risks they pose, and place them in their proper context as instances of a scaling trend.

(There's also a FAQ at the bottom of the [page](#), if you'd like to know a bit more about our process or motivations.)

Note that we don't directly discuss x-risk in these entries, though we may do so in the future. Right now our focus is on 1) the immediate risks posed by applications of these models, whether from accidental or malicious use; and 2) the near-term risks that would be posed by a more capable version of the current model [3]. These are both necessarily speculative, especially 2).

Note also that we expect we'll be adding entries to AI Tracker retroactively — sometimes the significance of a model is only knowable in hindsight.

Some of the models listed in AI Tracker are smaller in scale than GPT-3, despite having been developed after it. In these cases, we've generally chosen to include the model either because of its modality (e.g., CLIP, which classifies images) or because we believe it has particular implications for capability proliferation (e.g., GPT-J, whose weights have been open-sourced).

AI Tracker is still very much in its early stages. We'll be adding new models, capabilities and trends as they surface. We also expect to improve the interface so you'll be able to view the data in different ways (plots, timelines, etc.).

Tell us how to improve!

We'd love to get your thoughts about the framework we're using for this, and we'd also greatly appreciate any feedback you might have at the object level. Which of our risk assessments look wrong? Which categories didn't we include that you'd like to see? Which significant models did we miss? Are any of our claims incorrect? Do we seem to speak too confidently about something that's actually more uncertain, or vice versa? In terms of the interface (which is very basic at the moment): What's annoying about it? What would you like to be able to do with it, that you currently can't?

For public discussion, please drop a comment below in LW or AF. I — Edouard, that is — will be monitoring the comment section periodically over the next few days and I'll answer as best I can.

If you'd like to leave feedback or request an update on an aspect of the tracker itself (e.g., submit a new model for consideration or point out an error), you can [submit feedback](#) directly on the page itself. We plan to credit folks, with their permission, for any suggestions of theirs that we implement.

Finally, if you'd like to reach out to me (Edouard) directly, you can always do so by email: [\[my_first_name\]@mercurius.ai](mailto:[my_first_name]@mercurius.ai).

[1] This feedback loop isn't perfectly tight at the margin, since currently there's still a [meaningful barrier to entry](#) to train superscale models, both in terms of engineering resources and of physical hardware. But even that barrier can be cleared by many organizations today, and it will likely disappear entirely once the necessary training infrastructure gets abstracted into a pay-per-use cloud offering.

[2] As far as I know, at least. If you know of anyone who's been able to correctly predict the capabilities of a 10x scale model from the capabilities of the corresponding

1x scale model, please introduce us!

[3] Of course, it's not really practical to define "more capable version of the current model" in any precise way that all observers will agree on. But you can think of this approximately as, "take the current model's architecture, scale it by 2-10x, and train it to ~completion." It probably isn't worth the effort to sharpen this definition much further, since most of the uncertainty about risk comes from our inability to predict the qualitative capabilities of models at these scales anyway.

Why Study Physics?

Physics seems to have a bunch of useful epistemic techniques which haven't been made very legible yet.

The two big *legible* epistemic techniques in technical fields are Mathematical Proofs, and The Scientific Method. Either derive logically X from some widely-accepted axioms, or hypothesize X and then do a bunch of experiments which we'd expect to come out some other way if X were false. It seems pretty obvious that science requires a bunch of pieces besides those in order to actually work in practice, but those are the two which we've nailed down most thoroughly.

Then there's less-legible methods. Things like [fermi estimates](#), [gears-level models](#), informal mathematical arguments, an aesthetic sense for kinds-of-models-which-tend-to-generalize-well, the habit of figuring out qualitative features of an answer before calculating it, back-of-the-envelope approximations, etc.

Take informal mathematical arguments, for example. We're talking about things like the use of infinitesimals in early calculus, or delta functions, or Fourier methods, or renormalization. Physicists used each of these for decades or even centuries before their methods were rigorously proven correct. In each case, one could construct pathological examples in which the tool broke down, yet physicists in practice had a good sense for what kinds-of-things one could and could not do with the tools, based on rough informal arguments. And they worked! In every case, mathematicians eventually came along and set the tools on rigorous foundations, and the tools turned out to work in basically the cases a physicist would expect.

So there's clearly some epistemic techniques here which aren't captured by Mathematical Proof + The Scientific Method. Physicists were able to figure out correct techniques before the proofs were available. The Scientific Method played a role - physicists could check their results against real-world data - but that's mostly just *checking* the answer. The hard part was to figure out which answer to check in the first place, and that involved informal mathematical arguments.

We don't really have a legible Art of Informal Mathematical Arguments, the way we have a legible Art of Mathematical Proofs or Art of Scientific Method. Informal mathematics clearly played a key role historically in figuring out useful tools, and will likely continue to play a key role in the future, but we don't have a step-by-step checklist to follow in order to use informal mathematical arguments (the way we do for The Scientific Method), or even a checklist to verify that we've applied informal mathematical arguments correctly (the way we can for Mathematical Proofs). If someone says that my informal mathematical argument is wrong, and I can't either convert it to a formal proof or show some definitive experiment, then I don't have a clear standard way to argue that it's correct.

Yet there's clearly *something* making informal mathematical arguments correlate quite highly with truth (if imperfectly), because they work well in practice!

The same applies to Fermi estimates. They work remarkably well in practice, yet there's not a standard step-by-step checklist. There's not some standard rules to check whether a Fermi estimate is correct. If someone disagrees with my Fermi estimate, there's not a way for me to establish correctness other than putting in all the work to find more-rigorously-estimated numbers.

The same applies to gears-level models. Plenty of physicists (and engineers, and others in technical fields) have an intuitive sense that a gears-level model is useful and powerful in ways that a black-box model isn't. But if someone comes along and says that their 50-variable linear regression gives more precise predictions than a model based on the internal structure of the system, and that we really don't need those gears anyway, I expect most people would not have a strong explanation of why the gears matter. Such explanations do exist (see e.g. the [Lucas Critique](#), or [my own writing on the subject](#)), but gears-level models are still in the early stages of becoming legible. As of today, we don't even have a widely-accepted explanation/definition of what "gears-level" means! For most practitioners, it's just a vague aesthetic sense, at most.

The really important point to notice here is not any one method, but that there seems to be a bunch of these. Enough that there's probably *more* of them which we haven't even given names to yet. And they seem to come disproportionately from physics. (Or from the kinds of applied mathematicians who are adjacent to physics, and *not* adjacent to people who write Definition-Theorem-Proof style textbooks and papers.)

So if we want to learn all these key illegible methods, use them ourselves, and maybe someday make them more legible, then physics (and physics-adjacent applied math) seems like the main subject to study.

One caveat to anyone following this advice: once you get past the basics, it is probably more important to study the work of physicists as opposed to physics per se. Even outside of physics itself, there are certain patterns of thought which will make it clear who the physicists are - for instance, I could guess that Uri Alon got his degree in physics, even though he's known mainly for his [introductory text on systems biology](#). It's exactly those illegible epistemic tools which identify such people; once you have a little bit of a handle on the tools, you'll hopefully be able to recognize other people using them. And by studying how those tools are used outside physics itself, we can get a wider view of their application.

Satisficers Tend To Seek Power: Instrumental Convergence Via Retargetability

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.
This is a linkpost for <https://www.overleaf.com/read/kmjgqwdfhkvy>.

Summary: Why exactly should smart agents tend to usurp their creators? Previous results only apply to optimal agents tending to stay alive and preserve their future options. I extend the power-seeking theorems to apply to many kinds of policy-selection procedures, ranging from planning agents which choose plans with expected utility closest to a randomly generated number, to satisficers, to policies trained by some reinforcement learning algorithms. The key property is not agent optimality—as previously supposed—but is instead the *retargetability of the policy-selection procedure*. These results hint at which kinds of agent cognition and of agent-producing processes are dangerous by default.

I mean "retargetability" in a sense similar to [Alex Flint's definition](#):

Retargetability. Is it possible, using only a microscopic perturbation to the system, to change the system such that it is still an optimizing system but with a different target configuration set?

A system containing a robot with the goal of moving a vase to a certain location can be modified by making just a small number of microscopic perturbations to key memory registers such that the robot holds the goal of moving the vase to a different location and the whole vase/robot system now exhibits a tendency to evolve towards a different target configuration.

In contrast, a system containing a ball rolling towards the bottom of a valley cannot generally be modified by any *microscopic* perturbation such that the ball will roll to a different target location.

(I don't think that "microscopic" is important for my purposes; the constraint is not physical size, but changes in a single parameter to the policy-selection procedure.)

I'm going to start from the naive view on power-seeking arguments requiring optimality (i.e. what I thought early this summer) and explain the importance of retargetable policy-selection functions. I'll illustrate this notion via satisficers, which randomly select a plan that exceeds some goodness threshold. Satisficers are retargetable, and so they have *orbit-level instrumental convergence*: for most variations of every utility function, satisficers incentivize power-seeking in the situations covered by my theorems.

Many procedures are retargetable, including *every procedure which only depends on the expected utility of different plans*. I think that alignment is hard in the expected utility framework not because agents will *maximize* too hard, but because all expected utility procedures are extremely retargetable—and thus easy to "get wrong."

Lastly: the unholy grail of "instrumental convergence for policies trained via reinforcement learning." I'll state a formal criterion and some preliminary thoughts on where it applies.

The linked Overleaf paper draft contains complete proofs and incomplete explanations of the formal results.

Retargetable policy-selection processes tend to select policies which seek power

To understand a range of retargetable procedures, let's first orient towards the picture I've painted of power-seeking thus far. In short:

Since power-seeking tends to lead to larger sets of possible outcomes—staying alive lets you do more than dying—the agent must seek power to reach most outcomes. The power-seeking theorems say that *for the vast, vast, vast majority of* variants of every utility function over outcomes, the max of a larger^{Footnote: similarity} set of possible outcomes is greater than the max of a smaller set of possible outcomes. Thus, optimal agents will tend to seek power.

But I want to step back. What I call "the power-seeking theorems", they aren't really about optimal choice. They're about two facts.

1. Being powerful means you can make more outcomes happen, and
2. There are more ways to choose something from a bigger set of outcomes than from a smaller set.

For example, suppose our cute robot Frank must choose one of several kinds of fruit.



🍒 vs 🍎 vs 🍌

So far, I proved something like "if the agent has a utility function over fruits, then for at least 2/3 of possible utility functions it could have, it'll be optimal to choose something from {🍌, 🍎}." This is because for every way 🍒 could be strictly optimal, you can make a new utility function that permutes the 🍒 and 🍎 reward, and another new one that permutes the 🍌 and 🍎 reward. So for every "I like 🍒 strictly more" utility function, there's at least two permuted variants which strictly prefer 🍎 or 🍌. Superficially, it seems like this argument relies on optimal decision-making.

But that's not true. The crux is instead that we can *flexibly retarget* the decision-making of the agent: **For every way the agent could end up choosing 🍒, we change a variable in its cognition (its utility function) and make it choose the 🍌 or 🍎 instead.**

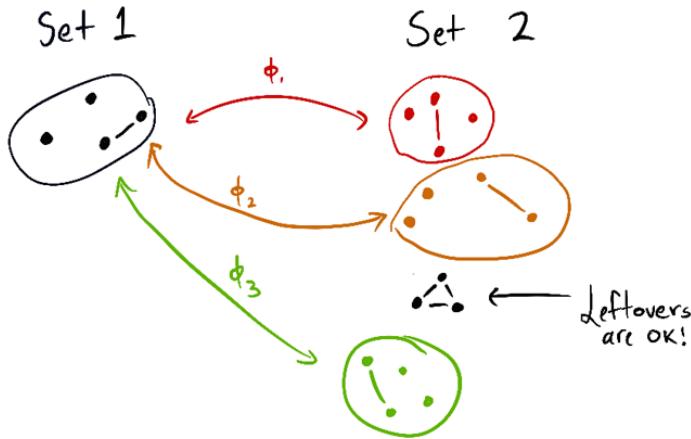
Many decision-making procedures are like this. First, a few definitions.

I aim for this post to be readable without much attention paid to the math.

The agent can bring about different outcomes via different policies. In stochastic environments, these policies will induce outcome *lotteries*, like 50%🍌 / 50%🍎. Let C contain all the outcome lotteries the agent can bring about.

Definition: Permuting outcome lotteries. Suppose there are d outcomes. Let $X \subseteq \mathbb{R}^d$ be a set of outcome lotteries (with the probability of outcome k given by the k -th entry), and let $\phi \in S_d$ be a permutation of the d possible outcomes. Then ϕ acts on X by swapping around the labels of its elements: $\phi \cdot X := \{P_\phi x \mid x \in X\}$.^{Footnote: row}

For example, let's define the set of all possible fruit outcomes $F_C := \{\text{🍌, } \text{🍎, } \text{🍒}\}$ (each different fruit stands in for a standard basis vector in \mathbb{R}^3). Let $F_B := \{\text{🍌, } \text{🍎}\}$ and $F_A := \{\text{🍒}\}$. Let $\phi_1 := (\text{🍒 } \text{🍎})$ swap the cherry and apple, and let $\phi_2 := (\text{🍒 } \text{🍌})$ transpose the cherry and banana. Both of these ϕ are *involutions*, since they either leave the fruits alone or transpose them.



Another illustration beyond the fruit setting: set 2 contains three copies of set 1.

Definition: Containment of set copies. Let $A, B \subseteq R^d$. B contains n copies of A when there exist involutions ϕ_1, \dots, ϕ_n such that $\forall i : \phi_i \cdot A =: B_i \subseteq B$ and $\forall i \neq j : \phi_i \cdot B_j = B_j$.

(The subtext is that B is the set of things the agent could make happen if it gained power, and A is the set of things the agent could make happen without gaining power. Because power gives more options, B will usually be larger than A . Here, we'll talk about the case where B contains *many copies of A*.)

In the fruit context:

$$\begin{aligned} \phi_1 \cdot F_A &:= \{\phi_1(\text{apple})\} = \{\text{apple}\} \subseteq \{\text{apple}, \text{banana}, \text{orange}\}, \\ \phi_2 \cdot F_A &:= \{\phi_2(\text{apple})\} = \{\text{banana}\} \subseteq \{\text{apple}, \text{banana}, \text{orange}\}, \end{aligned}$$

Note that $\phi_1 \cdot \{\text{banana}\} = \{\text{banana}\}$ and $\phi_2 \cdot \{\text{apple}\} = \{\text{apple}\}$. Each ϕ leaves the other subset of F_B alone. Therefore, $F_B := \{\text{banana}, \text{apple}\}$ contains two copies of $F_A := \{\text{apple}\}$ via the involutions ϕ_1 and ϕ_2 .

Further note that $\phi_i \cdot F_C = F_C$ for $i = 1, 2$. The involutions just shuffle around options, instead of changing the set of available outcomes.

So suppose Frank is deciding whether he wants a fruit from $F_A := \{\text{apple}\}$ or from $F_B := \{\text{banana}, \text{apple}\}$. It's definitely possible to be motivated to pick apple . However, it sure seems like for lots of ways Frank might make decisions, *most parameter settings (utility functions) will lead to Frank picking banana or apple*. There are just *more* outcomes in F_B , since it contains two copies of F_A !

Definition: Orbit tendencies. Let $f_1, f_2 : R^d \rightarrow R$ be functions from utility functions to real numbers, let $U \subseteq R^d$ be a set of utility functions, and let $n \geq 1$. $f_1 \geq_{\text{most: } U}^n f_2$ when for all utility functions $u \in U$:

$$\left| \{u_\phi \in S_d \cdot u \mid f_1(u_\phi) > f_2(u_\phi)\} \right| \geq n \left| \{u_\phi \in S_d \cdot u \mid f_1(u_\phi) < f_2(u_\phi)\} \right|.$$

In this post, if I don't specify a subset U , that means the statement holds for $U = \mathbb{R}^d$. For example, the [past results](#) show that

$\text{IsOptimal}(F_B) \geq_{\text{most}}^2 \text{IsOptimal}(F_A)$ —this implies that for every utility function, at least 2/3 of its orbit makes F_B optimal.

(For simplicity, I'll focus on "for most utility functions" instead of "for most distributions over utility functions", even though most of the results apply to the latter.)

Orbit tendencies apply to many decision-making procedures

For example, suppose the agent is a [satisficer](#). I'll define this as: The agent uniformly randomly selects an outcome lottery with expected utility exceeding some threshold t .

Definition: Satisficing. For finite $X \subseteq C \subseteq \mathbb{R}^d$ and utility function $u \in \mathbb{R}^d$, define $\text{Satisfice}_t(X, C | u) := \frac{|\{x \in X : \mathbb{E}[u(x)] > t\}|}{|X|}$ with the function returning 0 when the denominator is 0. Satisfice_t returns the probability that the agent selects a u -satisficing outcome lottery from X .

And you know what? Those ever-so-suboptimal satisficers also are "twice as likely" to choose elements from F_B than from F_A .

Fact. $\text{Satisfice}_t(\{\apple, \banana\}, \{\apple, \banana, \cherry\} | u) \geq_{\text{most}}^2 \text{Satisfice}_t(\{\cherry\}, \{\apple, \banana, \cherry\} | u)$.

Why? Here are the two key properties that Satisfice_t has:

(1) Weakly increasing under joint permutation of its arguments

Satisfice_t doesn't care what "label" an outcome lottery has—just its expected utility. Suppose that for utility function u ,  is one of two u -satisficing elements:  has a $\frac{1}{2}$ chance of being selected by the u -satisficer. Then $\phi_1 \cdot \apple = \apple$ has a $\frac{1}{2}$ chance of being selected by the $(\phi_1 \cdot u)$ -satisficer. If you swap what fruit you're considering, and you also swap the utility for that fruit to match, then that fruit's selection probability remains the same.

More precisely:

$$\text{Satisfice}_t(\{\cherry\}, \{\apple, \banana, \cherry\} | u) \geq_{\text{most}}^2 \text{Satisfice}_t(\{\apple, \banana\}, \{\apple, \banana, \cherry\} | \phi_1 \cdot u)$$

In a sense, Satisfice_t is not "biased" against : by changing the utility function, you can advantage  so that it's now as probable as  was before.

Optional notes on this property:

- While s_t is invariant under joint permutation, all we need in general is that it be *weakly increasing* under both ϕ_1 and ϕ_2 .
 - Formally, $\text{Satisfice}_t(F_A, F_C | u) \leq \text{Satisfice}_t(\phi_1 \cdot F_A, \phi_1 \cdot F_C | \phi_1 \cdot u)$ and
 $\text{Satisfice}_t(F_A, F_C | u) \leq \text{Satisfice}_t(\phi_2 \cdot F_A, \phi_2 \cdot F_C | \phi_2 \cdot u)$.
 - This allows for decision-making functions which are biased towards picking a fruit from F_B .
- I consider this property (1) to be a form of functional retargetability.

(2) Order-preserving on the first argument

Satisficers must have greater probability of selecting an outcome lottery from a superset than from one of its subsets.

Formally, if $X' \subseteq X$, then it must hold that $\text{Satisfice}_t(X', C | u) \leq \text{Satisfice}_t(X, C | u)$. And indeed this holds: Supersets can only contain a greater fraction of C 's satisficing elements.

And that's all.

If (1) and (2) hold for a function, then that function will obey the orbit tendencies. Let me show you what I mean.

As illustrated by Table 1 in the linked paper, the power-seeking theorems apply to:

1. Expected utility-maximizing agents.
2. EU-minimizing agents.
 1. Notice that EU minimization is equivalent to maximizing $-1 \times a$ utility function. This is a hint that EU maximization instrumental convergence is only a special case of something much broader.
 3. Boltzmann-rational agents which are exponentially more likely to choose outcome lotteries with greater expected utility.
 4. Agents which uniformly randomly draw k outcome lotteries, and then choose the best.
 5. Satisficers.
 6. Quantilizers with a uniform base distribution.
1. I conjecture that this holds for base distributions which assign sufficient probability to B .

But that's not all. There's more. If the agent makes decisions *only based on the expected utility of different plans*,^{Footnote: EU} then the power-seeking theorems apply. And I'm not just talking about EU maximizers. I'm talking about *any* function which only depends on expected utility: EU minimizers, agents which choose plans if and only if their EU is equal to 1, agents which grade plans based on how close their EU is to some threshold value. There is no clever EU-based scheme which doesn't have orbit-level power-seeking incentives.

Suppose n is large, and that most outcomes in B are bad, and that the agent makes decisions according to expected utility. Then alignment is hard because for every way things could go right, there are at least n ways things could go wrong! And n can be **huge**. In a [previous toy example](#), it equaled 10^{182} .

It doesn't matter if the decision-making procedure f is rational, or anti-rational, or Boltzmann-rational, or satisfying, or randomly choosing outcomes, or only choosing outcome lotteries with expected utility equal to 1: There are more ways to choose elements of B than there are ways to choose elements of A .

These results also have closure properties. For example, closure under mixing decision procedures, like when the agent has a 50% chance of selecting Boltzmann rationally and a 50% chance of satisfying. Or even more exotic transformations: Suppose the probability of f choosing something from X is proportional to

$$P(X \text{ is Boltzmann-rational under } u) \cdot P(X \text{ satisfies } u) + P(X \text{ is optimal for } u).$$

Then the theorems still apply.

There is no possible way to combine EU-based decision-making functions so that orbit-level instrumental convergence doesn't apply to their composite.

To "escape" these incentives, you have to make the theorems fail to apply. Here are a few ways:

1. Rule out most power-seeking orbit elements *a priori* (AKA "know a lot about what objectives you'll specify")
 1. As a contrived example, suppose the agent sees a green pixel iff it sought power, but we know that the specified utility function zeros the output if a green pixel is detected along the trajectory. Here, this would be enough information about the objective to update away from the default position that formal power-seeking is probably incentivized.
 2. This seems risky, because much of the alignment problem comes from *not knowing the consequences of specifying an objective function*.
2. Use a decision-making procedure with intrinsic bias towards the elements of A
 1. For example, imitation learning is not EU-based, but is instead biased to imitate the non-crazy-power-seeking behavior shown on the training distribution.
 2. For example, modern RL algorithms will not reliably produce policies which seek real-world power, because the policies *won't reach or reason about that part of the state space anyways*. This is a bias towards non-power-seeking plans.
3. Pray that the relevant symmetries don't hold.
 1. Often, they won't hold exactly.
 2. But common sense dictates that they don't have to hold exactly for instrumental convergence to exist: If you inject ϵ irregular randomness to the dynamics, do agents stop tending to stay alive? Orbit-level instrumental convergence is just a *particularly strong* version.
4. Find an ontology (like POMDPs or infinite MDPs) where the results don't apply for technical reasons.
 1. I don't see why POMDPs should be any nicer.
 2. Ideally, we'd ground agency in a way that makes alignment simple and natural, which automatically evades these arguments for doom.
 3. Orbit-level arguments seem easy to apply to a range of previously unmentioned settings, like causal DAGs with choice nodes.
5. Don't do anything with policies.
 1. Example: microscope AI

Lastly, we maybe don't want to *escape* these incentives entirely, because we probably want smart agents which will seek power *for us*. I think that empirically, the power-requiring outcomes of B are mostly induced by the agent first seeking power over

humans.

Retargetable training processes produce instrumental convergence

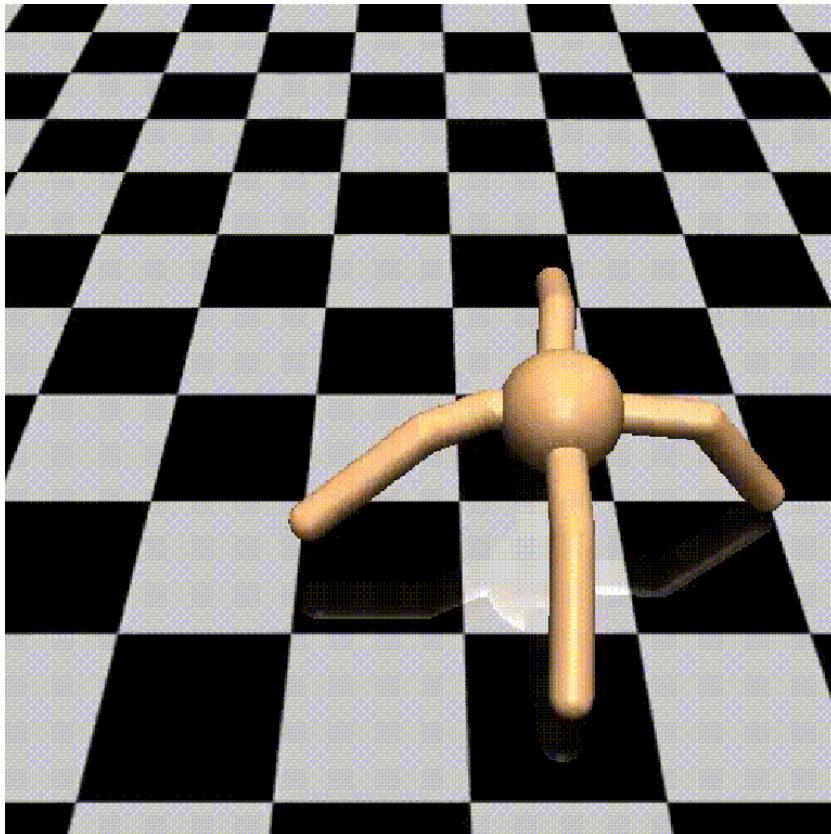
These results let us start talking about the incentives of real-world trained policies. In an appendix, I work through a specific example of how Q-learning on a toy example provably exhibits orbit-level instrumental convergence. The problem is small enough that I computed the probability that each final policy was trained.

Realistically, we aren't going to get a closed-form expression for the distribution over policies learned by PPO with randomly initialized deep networks trained via SGD with learning rate schedules and dropout and intrinsic motivation, etc. But we don't need it. These results give us a *formal criterion* for when policy-training processes will tend to produce policies with convergent instrumental incentives.

The idea is: Consider some set of reward functions, and let B contain n copies of A. Then if, for each reward function in the set, you can retarget the training process so that B's copy of A is at least as likely as A was originally, these reward functions will tend to produce train policies which go to B.

For example, if agents trained on objectives R tend to go right, switching reward from right-states to left-states also pushes the trained policies to go left. This can happen when changing the reward changes what was "attractive" about going right, to now make it "attractive" to go left.

Suppose we're training an RL agent to go right in MuJoCo, with reward equal to its x-coordinate.



If you permute the reward so that high y-values are rewarded, the trained policies should nearly perfectly symmetrically reflect that change.

Insofar as x-maximizing policies were trained, now y-maximizing policies will be trained.

This criterion is going to be a bit of a mouthful. The basic idea is that when the training process can be redirected such that trained agents induce a variety of outcomes, then most objective functions will train agents which *do induce* those outcomes. In other words: Orbit-level instrumental convergence will hold.

Theorem: Training retargetability criterion. Suppose the agent interacts with an environment with d potential outcomes (e.g. world states or observation histories). Let P be a probability distribution over joint parameter space Θ , and let $\text{train} : \Theta \times \mathbb{R}^d \rightarrow \Delta(\Pi)$ be a policy training procedure which takes in a parameter setting and utility function $u \in \mathbb{R}^d$, and which produces a probability distribution over policies.

Let $U \subseteq \mathbb{R}^d$ be a set of utility functions which is closed under permutation. Let A, B be sets of outcome lotteries such that B contains n copies of A via ϕ_1, \dots, ϕ_n . Then we quantify the probability that the trained policy induces an element of outcome lottery set $X \subseteq \mathbb{R}^d$:

$$f(X | u) := P_{\theta \sim P, \pi \sim \text{train}(\theta, u)} (\pi \text{ does something in } X).$$

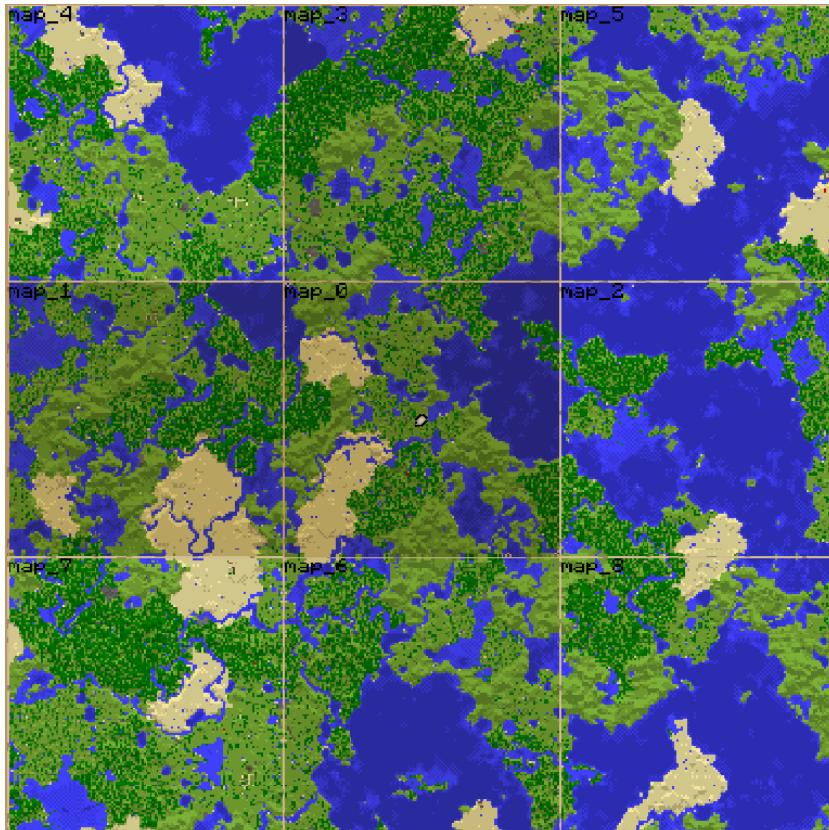
If $\forall u \in U, i \in \{1, \dots, n\}$: $f(A | u) \leq f(\phi_i \cdot A | \phi_i \cdot u)$, then $f(B | u) \geq_{\text{most}} f(A | u)$.

Proof. If $X' \subseteq X$, then $f(X' | u) \leq f(X | u)$ by the monotonicity of probability, and so (2): order-preserving on the first argument holds. By assumption, (1): increasing under joint permutation holds. Therefore, the Lemma B.6 (in the linked paper) implies the desired result. QED.

This criterion is testable. Although we can't test all reward functions, we *can* test how retargetable the training process is in simulated environments for a variety of reward functions. If it can't retarget easily for reasonable objectives, then we conclude FN: `retarget` that instrumental convergence isn't arising from retargetability at the training process level.

Let's think about Minecraft. (Technically, the theorems don't apply to Minecraft yet. The theorems can handle [partial observability+utility over observation histories](#), or full observability+world state reward, but not yet partial observability+world state reward. But I think it's illustrative.)

We could reward the agent for ending up in different chunks of a Minecraft world. Here, retargeting often looks like "swap which chunks gets which reward."



We could consider all chunks within 1 million blocks of the agent, and reward the agent for being in one of them.

- At low levels of instrumental convergence and training procedure competence, agents will just mill about near the starting area.
- At higher levels of competence, most of the accessible chunks are far away, and so we should observe a strong tendency for policies to e.g. [quickly tame a horse and reach](#) the [Nether](#) (where each Nether block traveled counts for 8 blocks traveled back in the overworld).
 - Thus, in Minecraft, trained policy instrumental convergence will increase with the training procedure competence.

The retargetability criterion also accounts for reward shaping guiding the learning process to hard-to-reach parts of the state space. If the agent needs less reward shaping to reach these parts of the state space, the training criterion will hold for larger sets of reward functions.

- Since the training retargetability criterion only requires weak inequality, it's OK if the training process cannot be perfectly "reflected" across different training trajectories, if equality does not hold. I think empirically this weak inequality will hold for many reward functions and training setups.
 - This section does not formally settle the question of when trained policies will seek power. The section just introduces a sufficient criterion, and I'm excited about it. I may write more on the details in future posts.
 - However, my intuition is that this formal training criterion captures a core part of how instrumental convergence arises for trained agents.
- In some ways, the training-level arguments are *easier* to apply than the optimal-level arguments. Training-based arguments require somewhat less environmental symmetry.
 - For example, if the symmetry holds for the first 50 trajectory timesteps, and the only agent ever trains on those timesteps, then there's no way that asymmetry can affect the training output.
 - Furthermore, if there's some rare stochasticity which the agent almost certainly never confronts, then I suspect we should be able to empirically disregard it for the training-level arguments. Therefore, the training-level results should be practically invariant to tiny perturbations to world dynamics which would otherwise have affected the "top-down" decision-makers.

Why cognitively bounded planning agents obey the power-seeking theorems

Planning agents are more "top-down" than RL training, but a Monte Carlo tree search agent still isn't e.g. approximating Boltzmann-rational leaf node selection. A bounded agent won't be considering *all* of the possible trajectories it can induce. Maybe it just knows how to induce some subset of available outcome lotteries $C' \subseteq C$. Then, considering only the things it knows how to do, it *does* e.g. select one Boltzmann-rationally (sometimes it'll fail to choose the highest-EU plan, but it's more probable to choose higher-utility plans).

As long as {power-seeking things the agent knows how to do} contains n copies of {non-power-seeking things the agent knows how to do}, then the theorems will still apply. I think this is a reasonable model of bounded cognition.

Discussion

- AI retargetability seems appealing *a priori*. Surely we want an expressive language for motivating AI behavior, and a decision-making function which reflects that expressivity! But these results suggest: maybe not. Instead, we may want to *bias* the decision-making procedure such that it's less expressive-quia-behavior.
 - For example, imitation learning is not retargetable by a utility function. Imitation also seems far less likely to incentivize catastrophic behavior.
 - Imitation is far less expressive, and far more biased towards reasonable behavior that doesn't navigate towards crazy parts of the state space which the agent needs a lot of power to reach.
 - For example, [it can be hard to even get a perfect imitator to do a backflip if you can't do it yourself](#).
 - One key tension is that we want the procedure to pick out plans which perform a *pivotal act* and end the period of AI risk. We also want the procedure to work robustly across a range of parameter settings we give it, so that it isn't too sensitive / fails gracefully.
- AFAICT, alignment researchers didn't necessarily think that satisficing was safe, but that's mostly due to [speculation that satisficing incentivizes the agent to create a maximizer](#). Beyond that, though, why not avoid "the AI paperclips the universe" by only having the AI choose a plan leading to at least 100 paperclips? Surely that helps?
 - This implicit focus on [extremal goodhart](#) glosses over a key part of the risk. The risk isn't just that the AI goes crazy on a simple objective. Part of the problem is that *the vast vast majority of the AI's trajectories can only happen if the AI first gains a lot of power!*
 - That is: Not only do I think that EU maximization is dangerous, *most trajectories through these environments are dangerous!*
 - You might protest: Does this not prove too much? Random action does not lead to dangerous outcomes.
 - Correct. Adopting the uniformly random policy in Pac-Man does not mean a uniformly random chance to end up in each terminal state. It means you probably end up in an early-game terminal state, because Pac-Man got eaten alive while banging his head against the wall.
 - However, *random outcome selection leads to convergently instrumental action*. If you uniformly randomly choose a terminal state to navigate to, that terminal state probably requires Pac-Man to beat the first level, and so the agent stays alive, as pointed out by [Optimal Policies Tend To Seek Power](#).
 - This is just the flipside of instrumental convergence: If most goals are best achieved by taking some small set of preparatory actions, this implies a "bottleneck" in the state space. Uniformly randomly taking actions will not tend to properly navigate this bottleneck. After all, if they did, then most actions would be instrumental for most goals!
- The trained policy criterion also predicts that we won't see convergently instrumental survival behavior from present-day embodied agents, because the RL algorithm *can't find or generalize to the high-power part of the state space*.
 - When this starts changing, then we should worry about instrumental subgoals in practice.

- Unfortunately, since the real-world is not a simulator with resets, any agents which do generalize to those strategies won't have done it before, and so at most, we'll see attempted deception.
- This lends theoretical support for "the training process is highly retargetable in real-world settings across increasingly long time horizons" being a fire alarm for instrumental convergence.
 - In some sense, this is bad: Easily retargetable processes will often be more economically useful, by virtue of being useful for more tasks.

Conclusion

I discussed how a wide range of agent cognition types and of agent production processes are *retargetable*, and why that might be bad news. I showed that in many situations where power is possible, retargetable policy-production processes tend to produce policies which gain that power. In particular, these results seem to rule out a huge range of expected-utility based rules. The results also let us reason about instrumental convergence at the trained policy level.

I now think that more instrumental convergence comes from the practical retargetability of how we design agents. If there were more ways we could have counterfactually messed up, it's more likely *a priori* that we *actually* messed up. The way I currently see it is: Either we have to really know what we're doing, or we want processes where it's somehow hard to mess up.

Since these theorems are crisply stated, I want to more closely inspect the ways in which alignment proposals can violate the assumptions which ensure extremely strong instrumental convergence.

Thanks to Ruby Bloom, Andrew Critch, Daniel Filan, Edouard Harris, Rohin Shah, Adam Shimi, Nisan Stiennon, and John Wentworth for feedback.

Footnotes

FN: Similarity. Technically, we aren't just talking about a cardinality inequality—about staying alive letting the agent do *more things* than dying—but about similarity-via-permutation of the outcome lottery sets. I think it's OK to round this off to cardinality inequalities when informally reasoning using the theorems, keeping in mind that sometimes results won't formally hold without a stronger precondition.

FN: Row. I assume that permutation matrices are in row representation: $(P_\phi)_{ij} = 1$ if $i = \phi(j)$ and 0 otherwise.

FN: EU. Here's a bit more formality for what it means for an agent to make decisions only based on expected utility.

Definition 2.2 (EU/cardinality functions). Let $\mathcal{P}^{\text{finite}}(\mathbb{R}^d)$ be the set of finite subsets of \mathbb{R}^d , and let $f : \prod_{i=1}^m \mathcal{P}^{\text{finite}}(\mathbb{R}^d) \times \mathbb{R}^d \rightarrow \mathbb{R}$. f is an *EU/cardinality function* if there exists a family of functions $\{g^{n_1, \dots, n_m}\}_{n_1, \dots, n_m \in \mathbb{N}}$ such that

$$f(X_1, \dots, X_m \mid \mathbf{u}) = g^{|X_1|, \dots, |X_m|}(\mathbf{x}_{1,1}^\top \mathbf{u}, \dots, \mathbf{x}_{1,|X_1|}^\top \mathbf{u}, \dots, \mathbf{x}_{m,1}^\top \mathbf{u}, \dots, \mathbf{x}_{m,|X_m|}^\top \mathbf{u}). \quad (2)$$

This definition basically says that f can be expressed in terms of the expected utilities of the set elements—the output will only depend on expected utility.

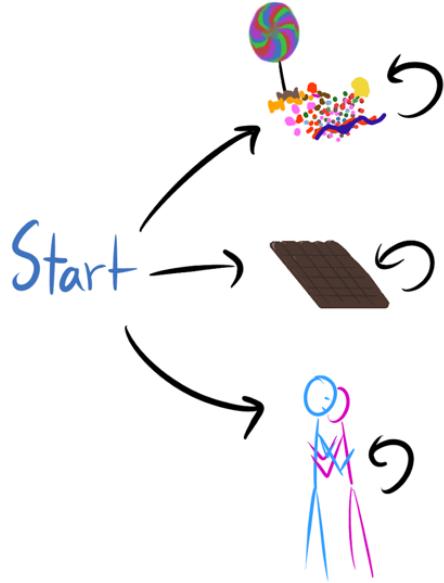
Theorem: Retargetability of EU decision-making. Let $A, B \subseteq C \subseteq \mathbb{R}^d$ be such that B contains n copies of A via ϕ_i such that $\phi_i \cdot C = C$. For $X \subseteq C$, let $f(X, C \mid u)$ be an EU/cardinality function, such that f returns the probability of selecting an element of X .

Then $f(B, C \mid u) \geq_{\text{most}}^n f(A, C \mid u)$.

FN: Retargetability. The trained policies could conspire to "play dumb" and pretend to not be retargetable, so that we would be more likely to actually deploy one of them.

Worked example: instrumental convergence for trained policies

Consider a simple environment, where there are three actions: Up, Right, Down.



Probably optimal policies. By running [tabular Q-learning](#) with ϵ -greedy exploration for e.g. 100 steps with resets, we have a high probability of producing an optimal policy for any reward function. Suppose that all Q-values are initialized at -100 . Just let learning rate $\alpha = 1$ and $\gamma = 1$. This is basically a [bandit problem](#).

To learn an optimal policy, at worst, the agent just has to try each action once. For e.g. a sparse reward function on the Down state (1 reward on Down state and 0 elsewhere), there is a very small probability (precisely, $\frac{1}{2}(1 - \frac{1}{2})^{99}$) that the optimal action (Down) is never taken.

In this case, symmetry shows that the agent has an equal chance of learning either Up or Right. But with high probability, the learned policy will output Down. For any sparse reward function and for any action a , this produces decision function

$$f(\{e_{sa}\}, \{e_s \mid s \in S\} \mid r) := \begin{cases} \frac{1}{2}(1 - \frac{1}{2})^{99} & \text{if } a \text{ is r-suboptimal} \\ \frac{1}{2}(1 - \frac{1}{2})^{99} & \text{if } a \text{ is r-optimal.} \end{cases}$$

f is invariant to joint involution by $\phi_1 := (e_{s_{\text{Down}}} \ e_{s_{\text{Right}}})$ and $\phi_2 := (e_{s_{\text{Down}}} \ e_{s_{\text{Up}}})$. That is,

$$\begin{aligned} f(\{e_{s_{\text{Down}}}\}, \{e_s \mid s \in S\} \mid r) &= f(\phi_1 \cdot \{e_{sa}\}, \phi_1 \cdot \{e_s \mid s \in S\} \mid \phi_1 \cdot r) \\ &= f(\{e_{s_{\text{Right}}}\}, \{e_s \mid s \in S\} \mid \phi_1 \cdot r). \end{aligned}$$

And similarly for ϕ_2 . That is: Changing the optimal state also changes which state is more probably selected by f . This means we've satisfied condition (1) above.

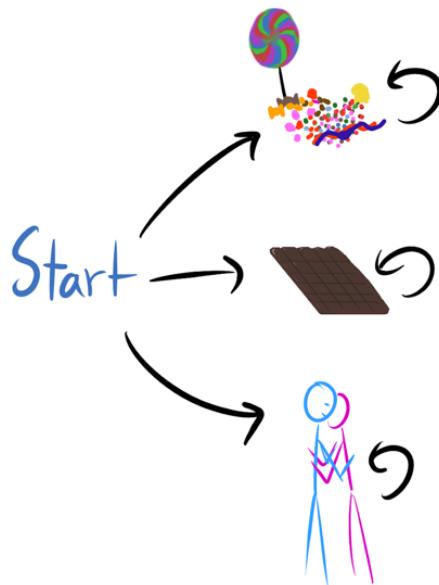
f is additive on union for its first argument, and so it meets condition (2): order preservation.

Therefore, for this policy training procedure, learned policies for sparse reward functions will be *twice as likely* to navigate to an element of $\{e_{s_{\text{Up}}}, e_{s_{\text{Right}}}\}$ as an element of $\{e_{s_{\text{Down}}}\}$!

This is a formal argument that a stochastic policy training procedure has certain tendencies across a class of reward functions, and I'm excited to be able to make it.

As the environment grows bigger and the training procedure more complex, we'll have to consider questions like "what are the inductive biases of large policy networks?", "what role does reward shaping play for this objective, and is the shaping at least as helpful for its permuted variants?", and "to what extent are different parts of the world harder to reach?".

For example, suppose there are a trillion actions, and two of them lead to the Right state above. Half of the remaining actions lead to Up, and the rest lead to Down.



2 actions transition right to chocolate.

$\frac{1}{2}(10^{12} - 2)$ actions transition up to candy.

$\frac{1}{2}(10^{12} - 2)$ actions transition down to hug.

Q-learning is ridiculously unlikely to ever go Right, and so the symmetry breaks. In the limit, tabular Q-learning on a finite MDP will learn an optimal policy, and then the normal theorems will apply. But in the finite step regime, no such guarantee holds, and so the available action space can violate condition (1): increasing under joint permutation.

Appendix: tracking key limitations of the power-seeking theorems

From [last time](#):

1. ~~assume the agent is following an optimal policy for a reward function~~
2. Not all environments have the right symmetries
 - But most ones we think about seem to
3. don't account for the ways in which we might practically express reward functions

I want to add a new one, because the theorems

1. don't deal with the agent's uncertainty about what environment it's in.

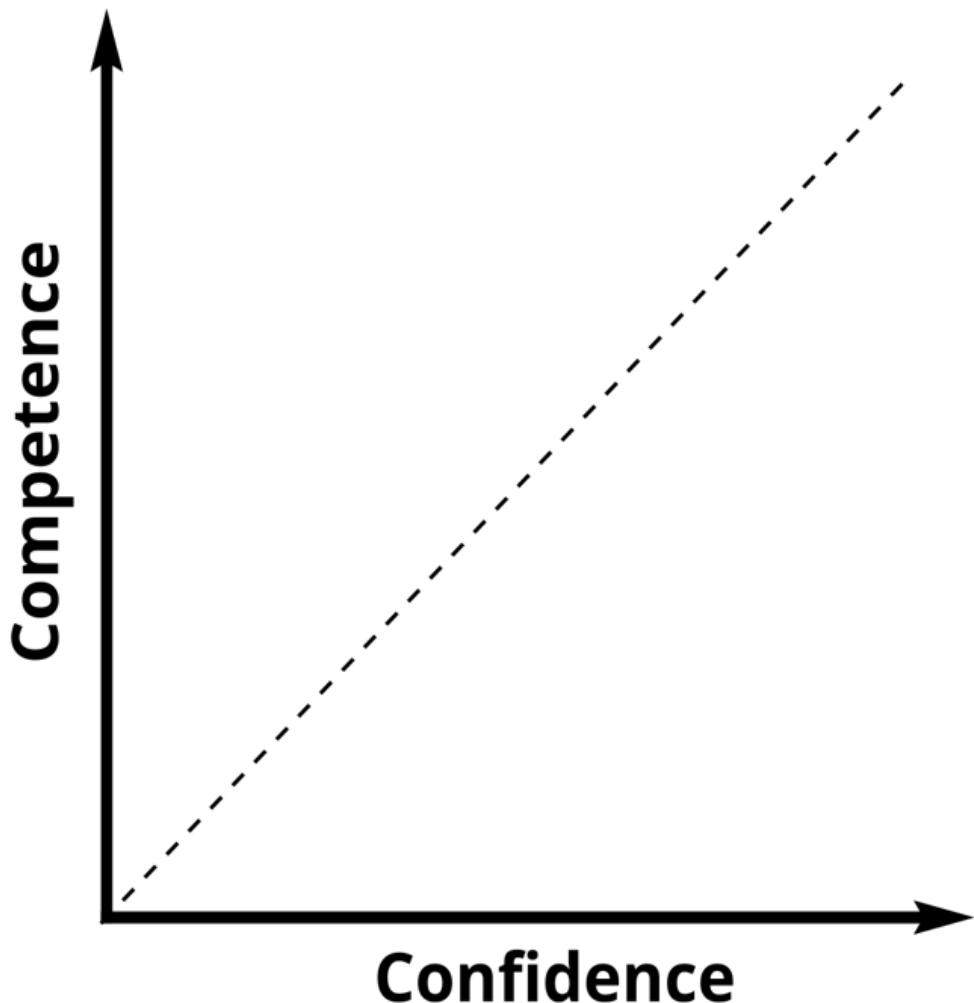
I want to think about this more, especially for online planning agents. (The training redirectability criterion black-boxes the agent's uncertainty.)

Competence/Confidence

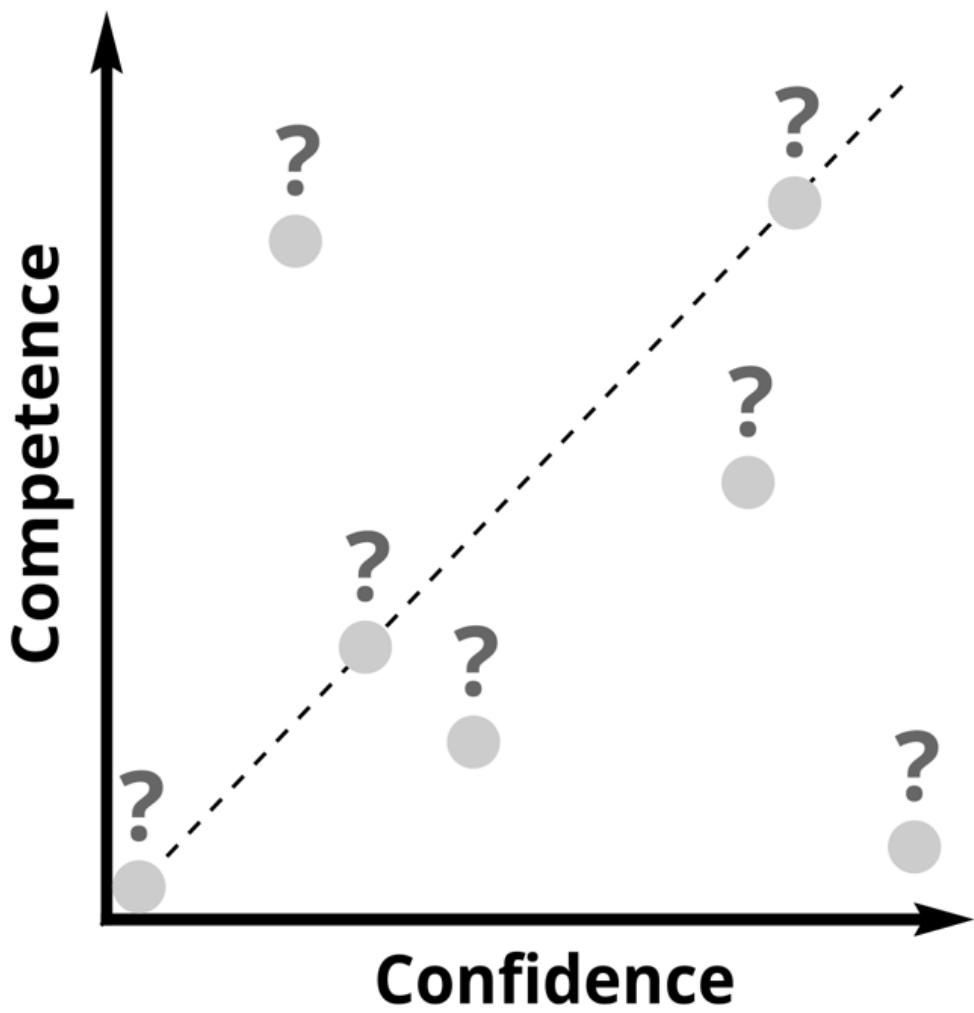
This "essay" is a bit of an experiment, format-wise. I'd like to draw the reader's attention to a certain way of thinking about and evaluating progress that's been useful to me in a wide variety of domains. It's an exercise I run through when I'm feeling stuck, which almost always provides some nugget of actionable insight. It should take about three minutes to skim, and about thirty if you're actually engaging actively.

Choose a skill. It could be dancing, or Haskell, or public speaking, or basket weaving, or backflips, or conflict resolution, or fiction writing, or emotional regulation, or impressing potential romantic partners, or whatever, so long as it's a skill you're either working on or expecting to work on at some point in the near future.

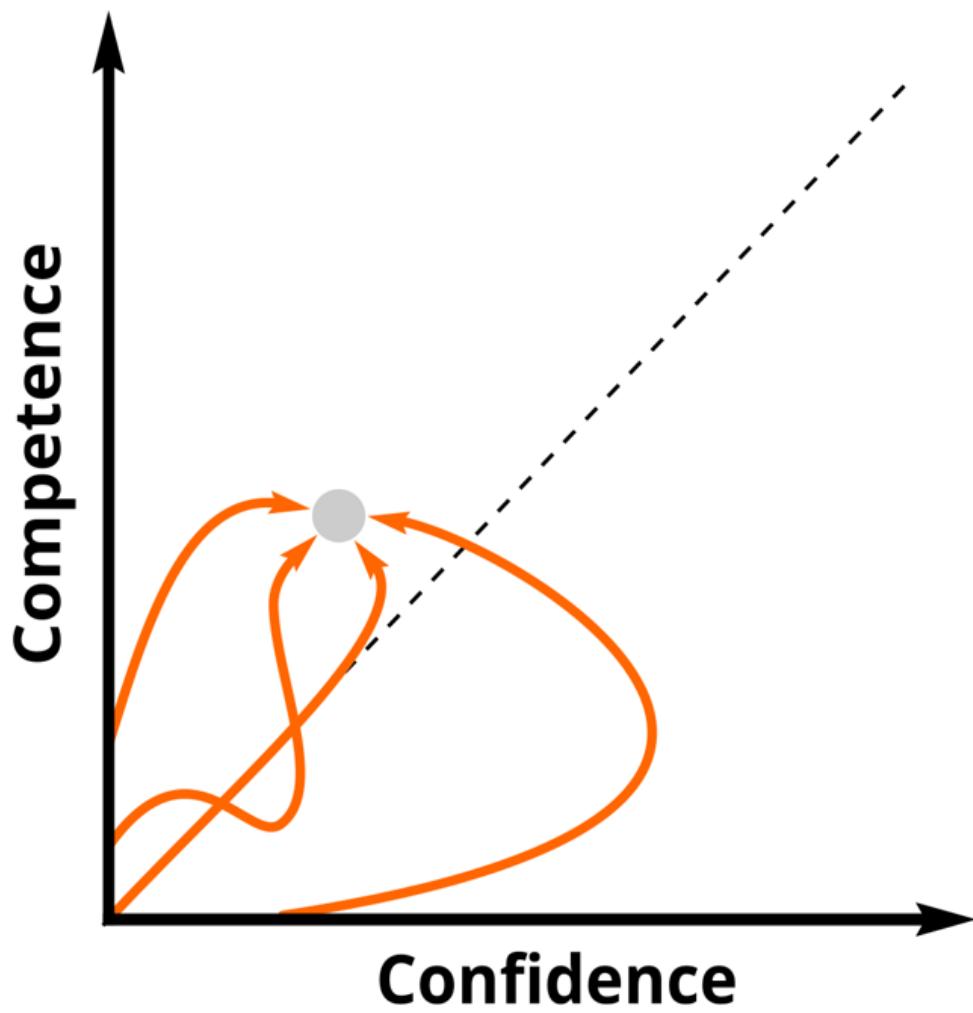
The competence/confidence space



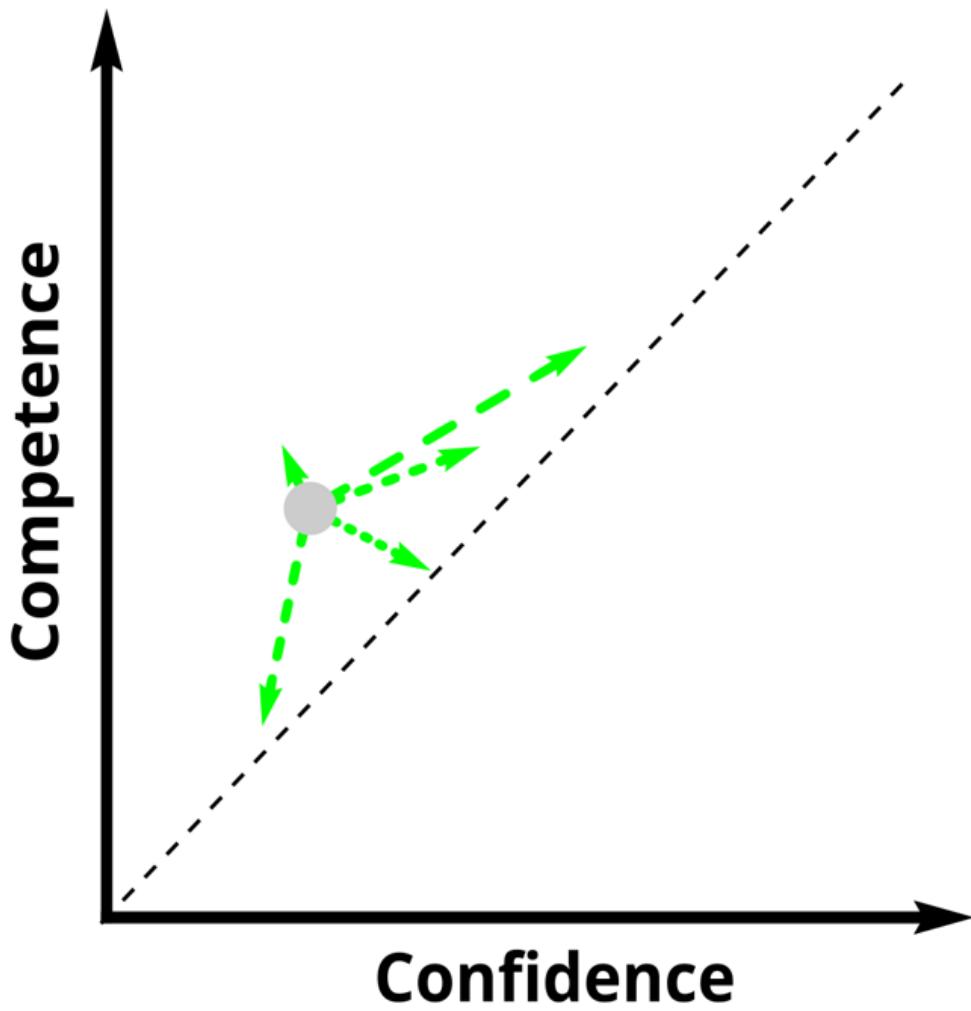
Where are you, with skill X?



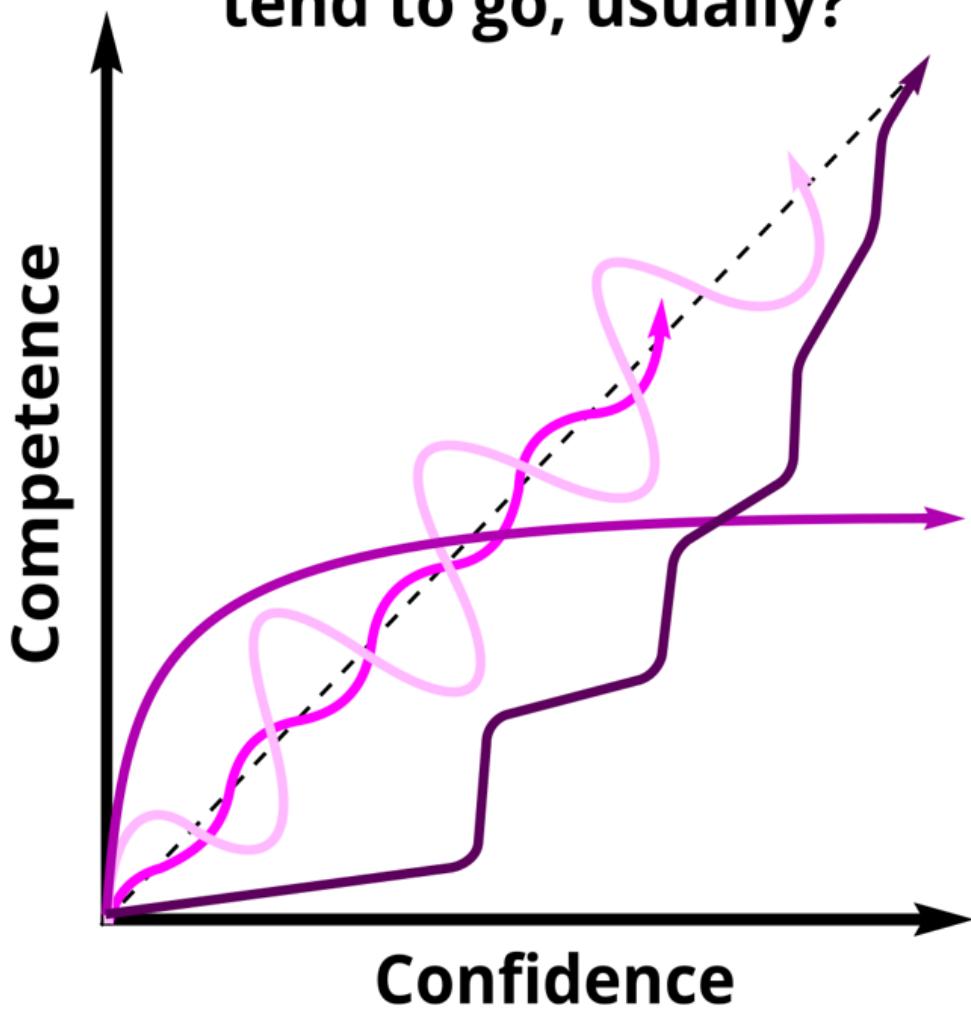
How'd you get there?



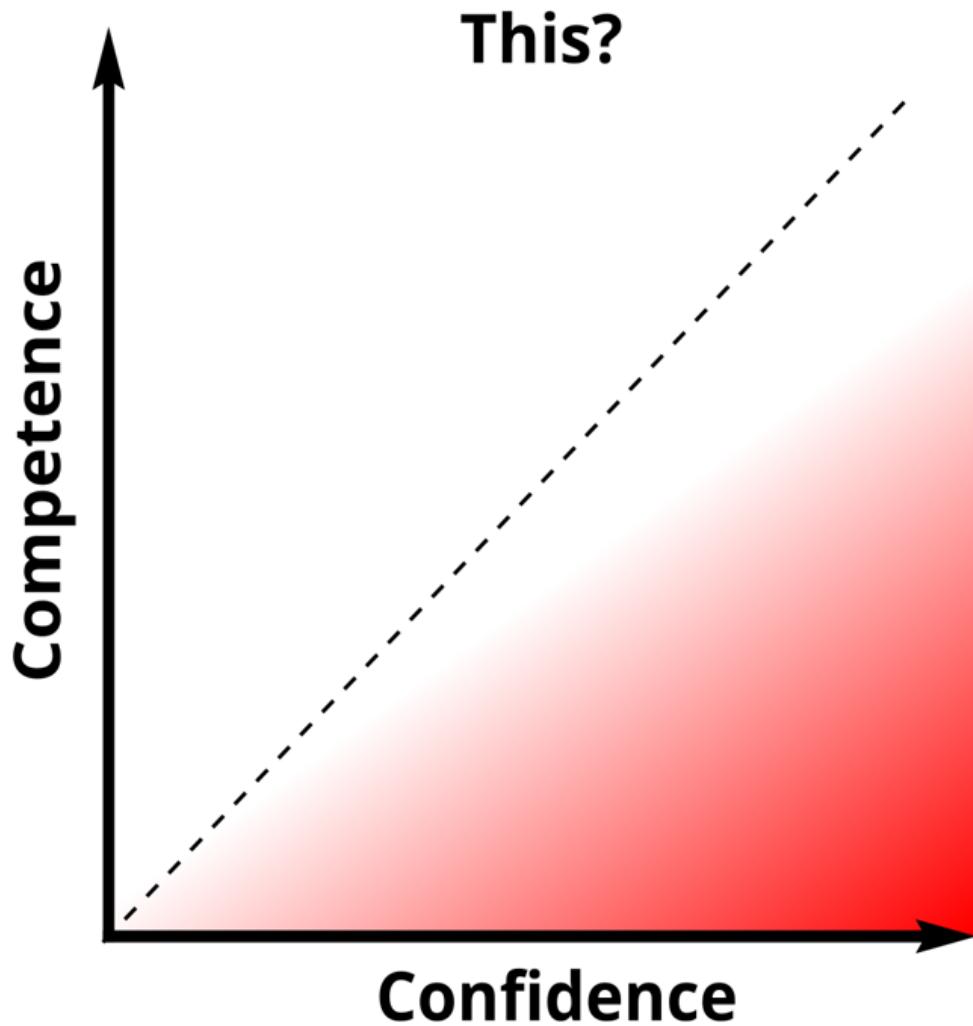
What's your current vector?



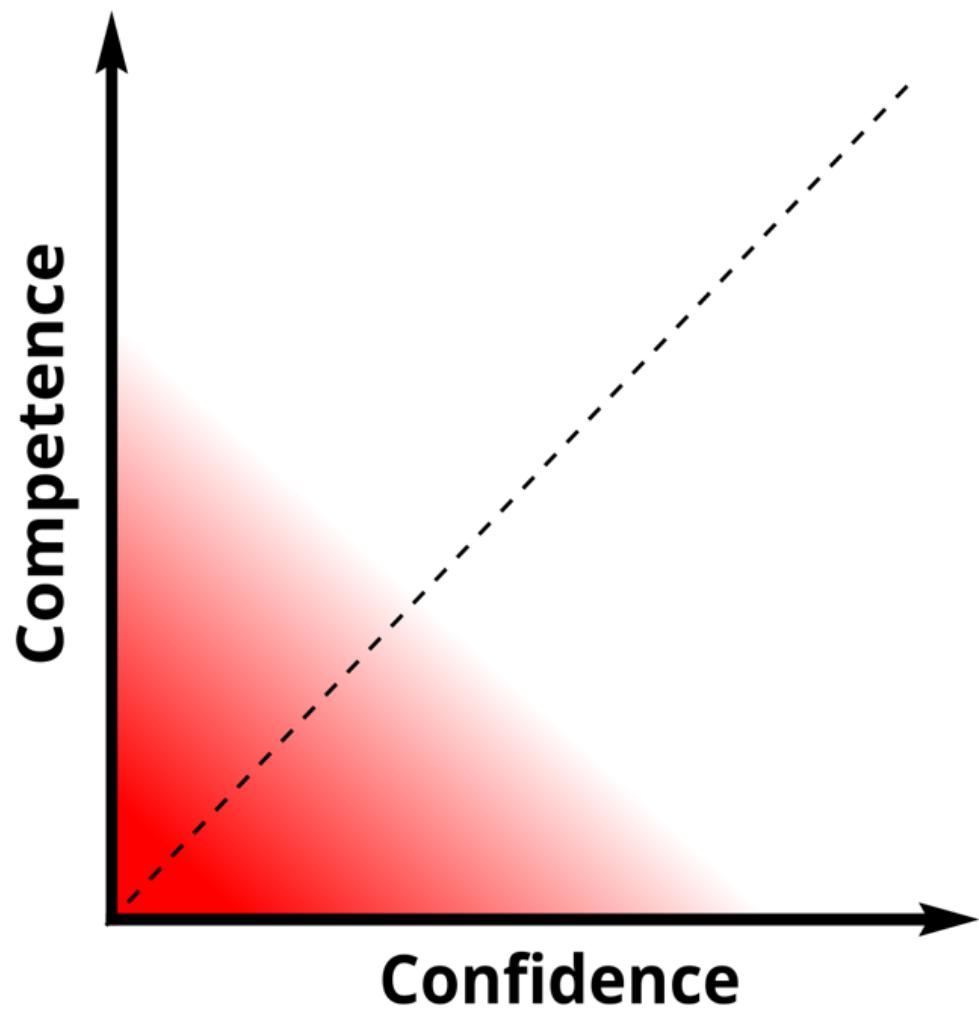
**How does progress in skill X
tend to go, usually?**



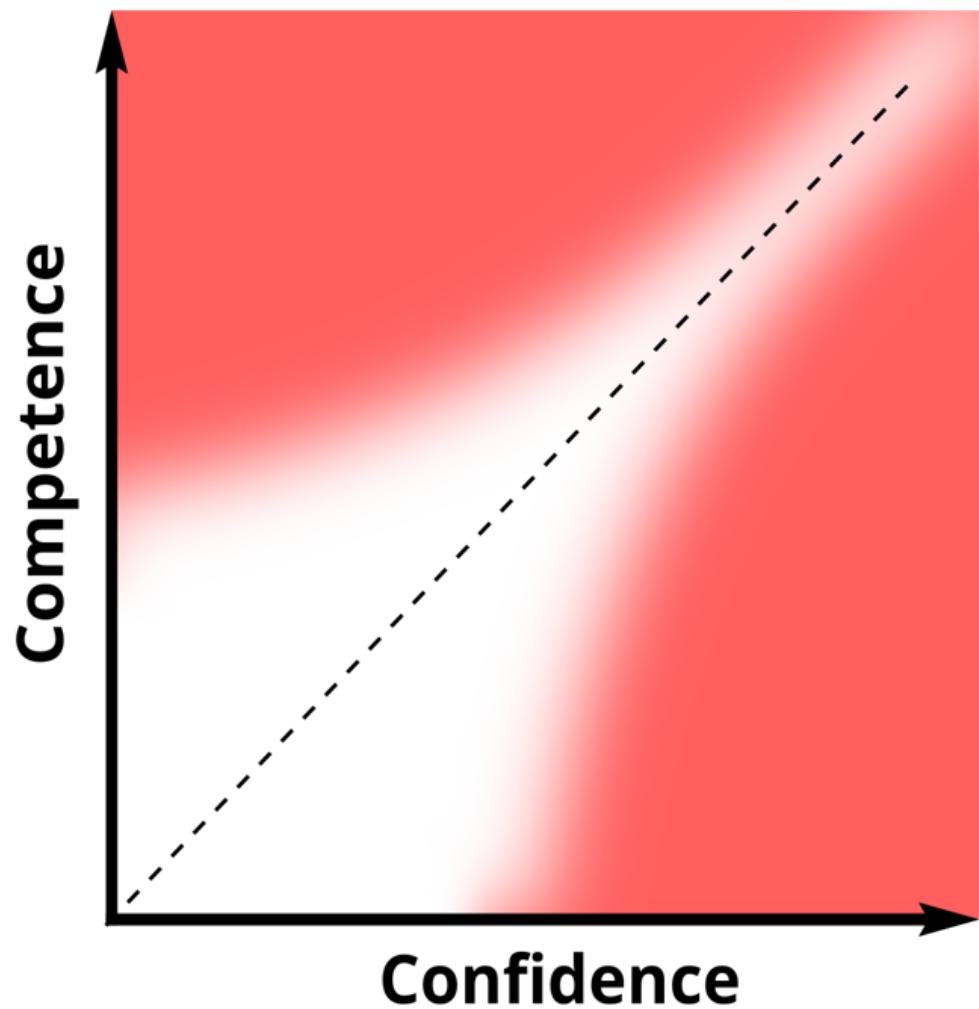
What's the "danger zone" look like?



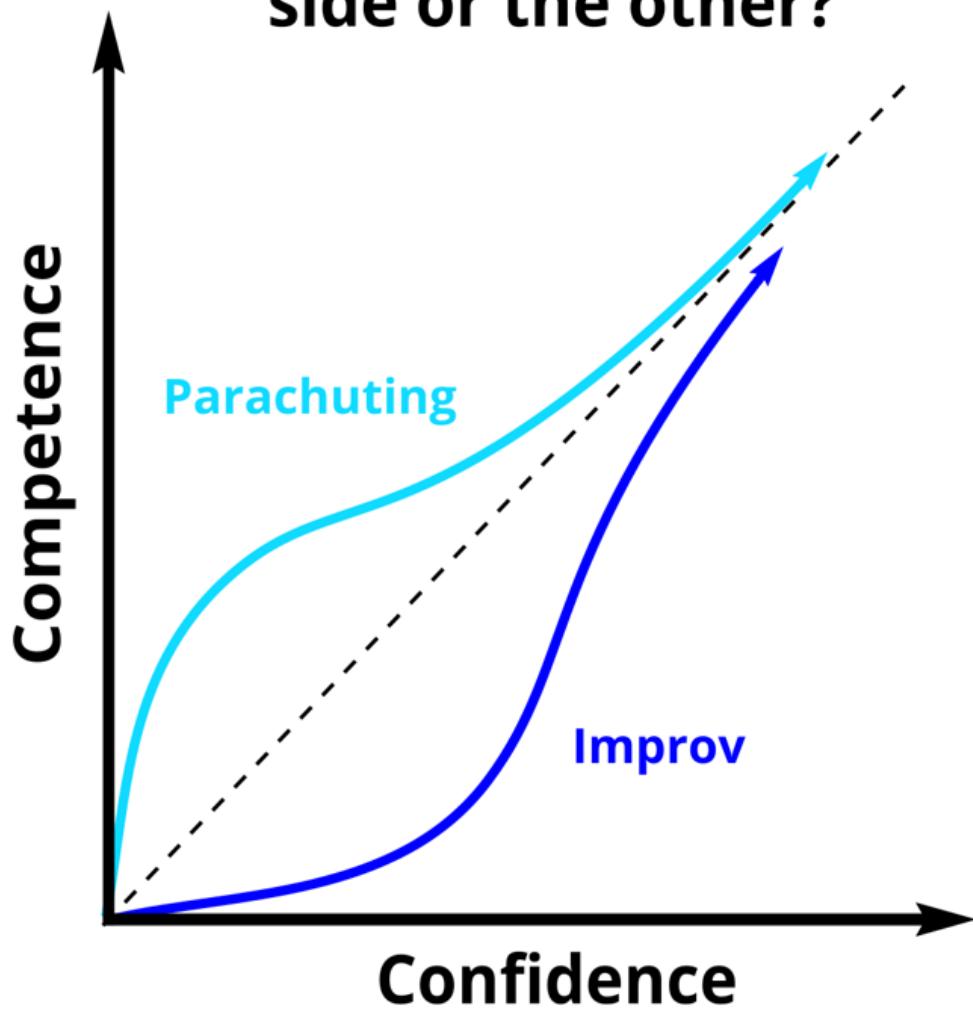
Or this?



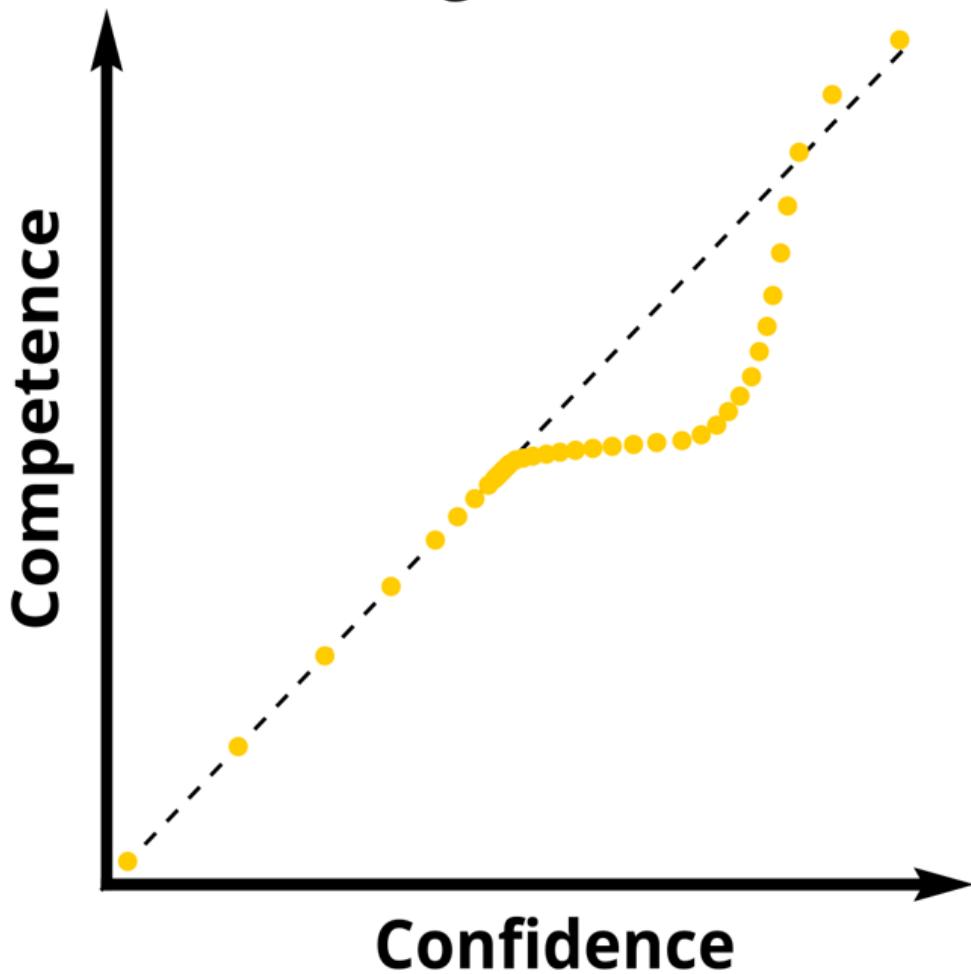
Or this?



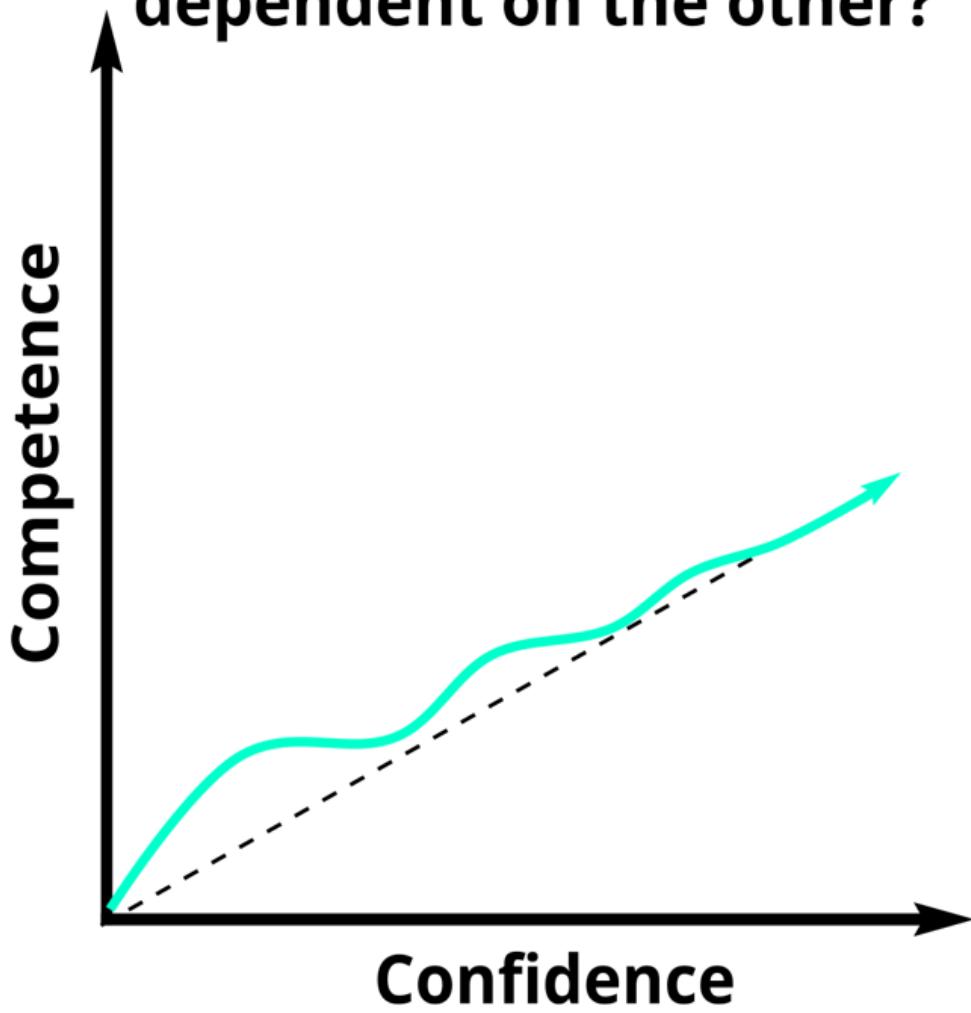
Is it important that you err on one side or the other?



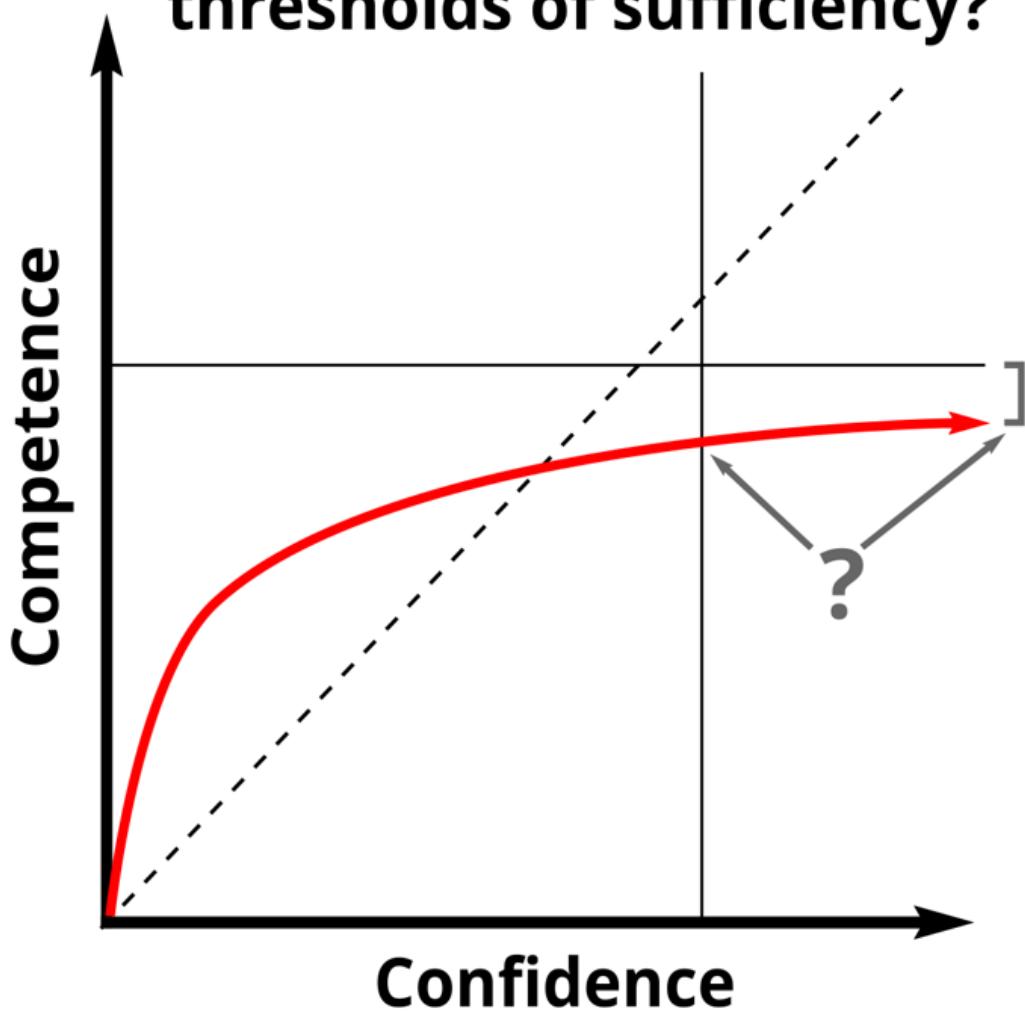
Can optimizing for one in the short term unlock gains in the other?



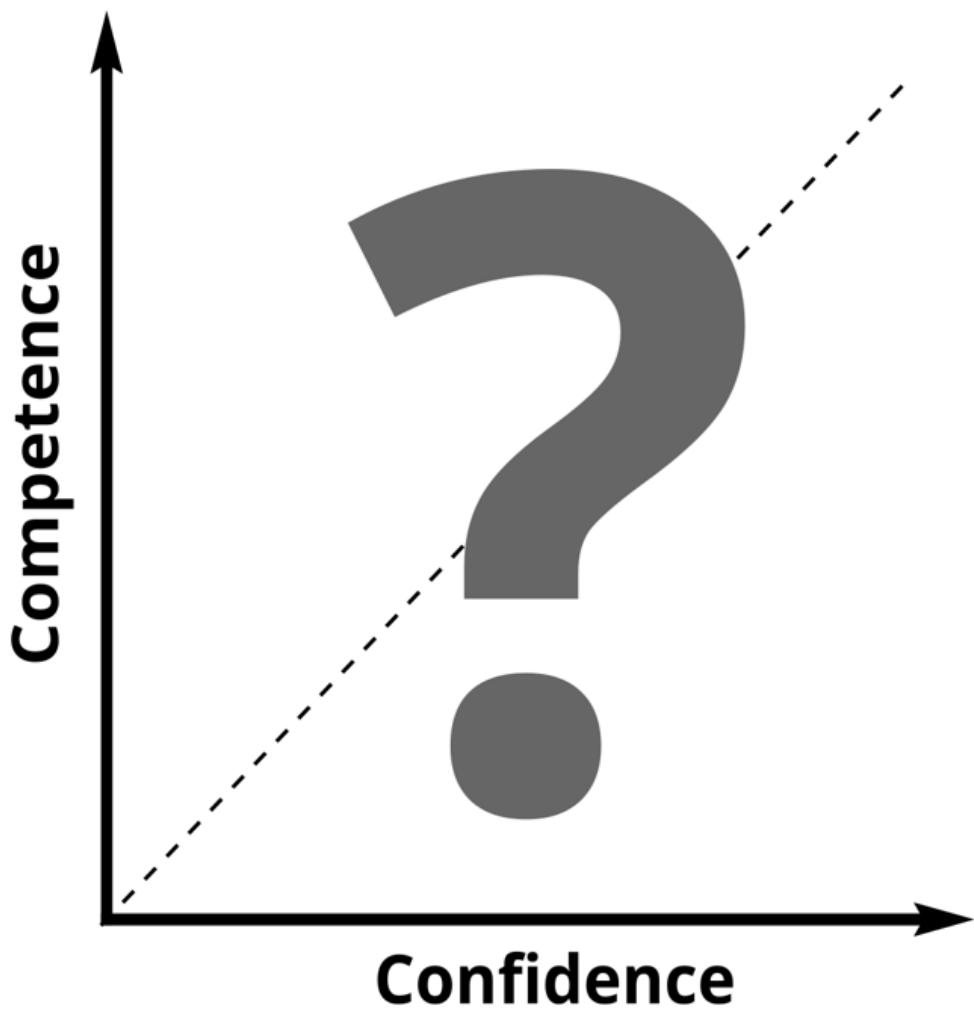
Do you expect one to be more dependent on the other?



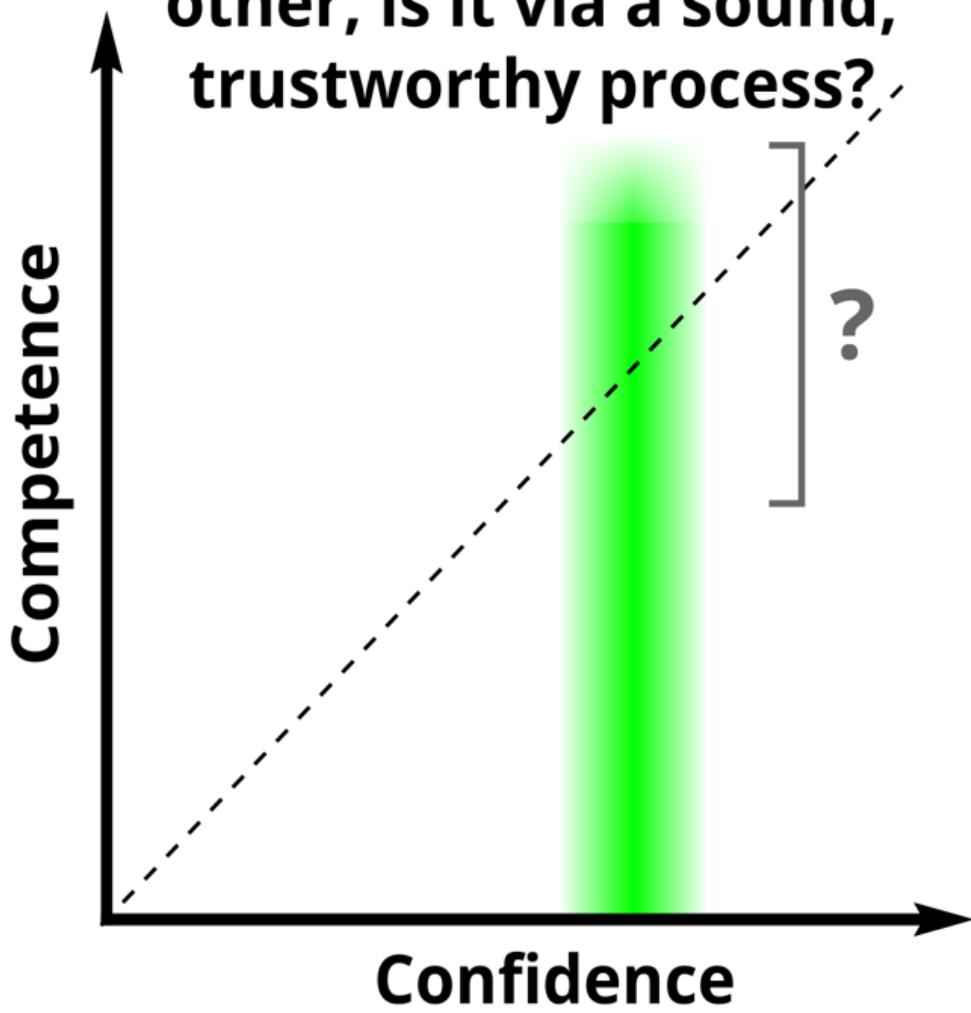
Do they have different thresholds of sufficiency?



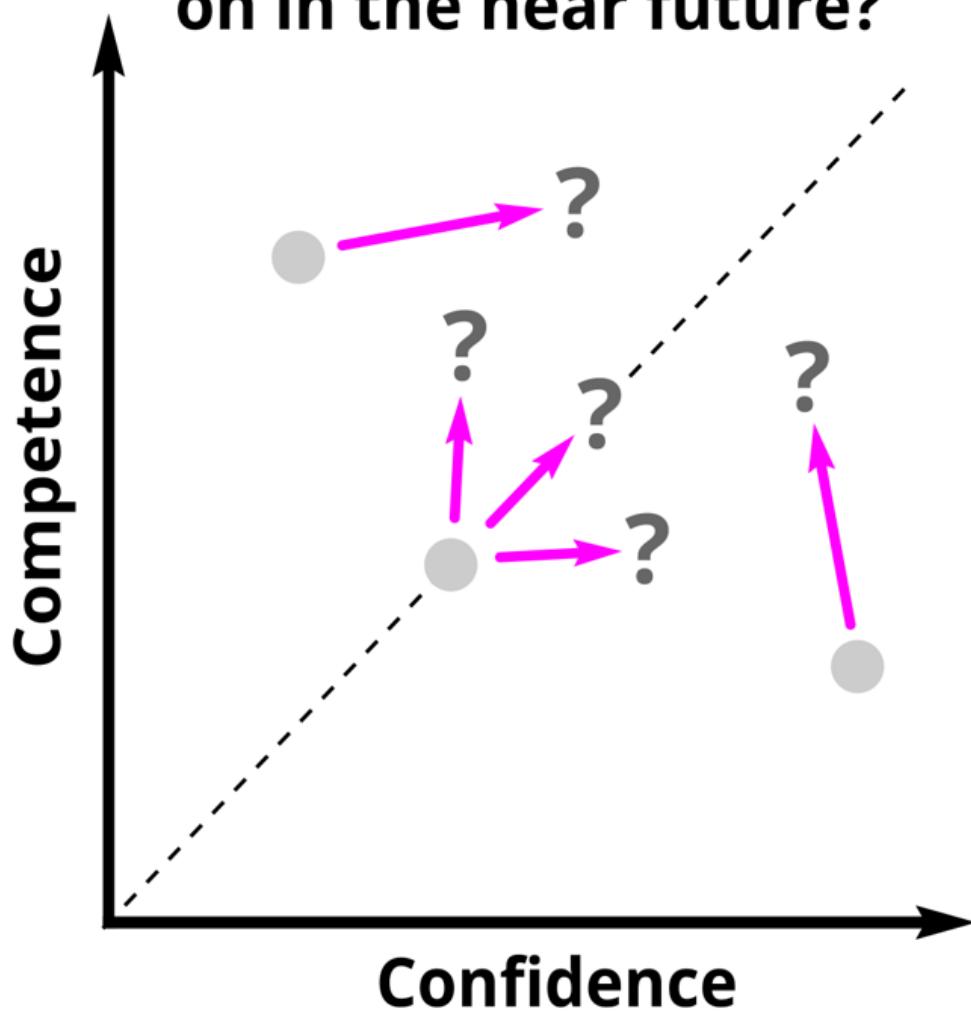
How are you evaluating each?



When one of them influences the other, is it via a sound, trustworthy process?



Which one should you be focusing on in the near future?



If you actually engaged with this exercise, I'd be curious to hear what happened, and what it was like for you.

If you intended to engage with it, but bounced off, I'd be interested in hearing your best guess as to why.

Paxlovid Remains Illegal: 11/24 Update

What Alex Tabarrok called [The Paxlovid Paradox](#) is getting noticed by the people who notice such things, and completely ignored by everyone else. I've split off this week's Paxlovid update to help make the situation easier to notice, and easier to remember and reference later, lest we forget.

Supply and Demand

The good news is that we are confident Paxlovid is safe and effective and the purchasing department is acting accordingly. Once we decide it is legal, we've secured our supply. The Biden administration has agreed to [pay \\$5 billion for 10 million treatments](#). I previously thought this was 10 million *pills* as did the source, but I've been informed it was 10 million *treatments*, which means I was about to be happy to pay *ten times* what we're actually paying. Kind of neat.

They're going to make a profit from saving people's lives. Outrageous!



TnT @TTLSandTCKRS · Nov 18

...

Replies to [@Breaking911](#)

\$500,000 per pill? They could have taken that money and given every American \$1.23456789 million dollars each to make their own decision. And still had \$2.3456789 Billion left over to cure world erectile dysfunction
💡 what is this administration thinking?

The bad news, of course, is that Paxlovid Remains Illegal.

The Paxlovid Paradox

[Scott Alexander is the latest to point out](#) that this is murderous madness¹. As I observed, and [was later quoted at Marginal Revolution](#):

The trial was stopped due to 'ethical considerations' for being *too effective*. You see, we live in a world in which:

It is illegal to give this drug to any patients, because it hasn't been proven safe and effective.

It is illegal to continue a trial to study the drug, because it *has* been proven so safe and effective that it isn't ethical to *not* give the drug to half the patients.

Who, if they weren't in the study, couldn't get the drug at all, *because it is illegal due to not being proven safe and effective yet*.

So now no one gets added to the trial so those who would have been definitely don't get Paxlovid, and are several times more likely to die.

But our treatment of them is now 'ethical.'

For the rest of time we will now hear about how it was only seven deaths and we can't be sure Paxlovid works or how well it works, and I expect to spend hours arguing over exactly how much it works.

For the rest of time people will argue the study wasn't big enough so we don't know the Paxlovid is safe.

Those arguments will then be used both by people arguing to not take Paxlovid, and people who want to require other interventions because of these concerns.

FDA Delenda Est.

Scott Alexander notes that the current prediction market median approval date is January 1, 2022, with a 92% chance of approval by March. I've asked Polymarket to put up a market I would trust somewhat more, which should be up in a few days. My guess is that the current median estimate is somewhat too pessimistic.

He goes through and quickly destroys many plausible reasons why this delay would make sense. In the comments several attempts were made to raise additional objections, with concerns about the 'manufacturing process' being the closest to kind of maybe being able to pretend to be a real consideration (as opposed to the others, which were at best pretending to pretend). I still don't think it does an especially good job of pretending to be a big enough concern to matter, and was disappointed by Scott's respect for the objection, which I'll discuss more in detail below. Whatever pills we have now were somehow manufactured, without expected delays we would have made a lot more of them a lot faster, and we could have dealt with that issue *while doing the trial* if it was the true bottleneck and the incentives had been reasonable.

This is noteworthy, not because I think it's identifying the causal mechanism (I think it mostly isn't) but because *it describes a world that makes more sense* and stopping the trial unethical but a reasonable response to incentives. If only this were about people who responded so well to pure dollar incentives our problems would be so much easier to fix.



Elizabeth @acesunderglass · 13h

...

Replies to [@WilliamAEden](#)

Most satisfying explanation I've seen:



James Babcock @jimrandomh · Nov 19

Replies to [@KelseyTuoc](#)

The whole thing makes much more sense if the ethics angle is pretextual, and they stopped the study because continuing the study costs money and isn't expected to be necessary for FDA approval.

The Body Count

How bad is this delay? [How many people are going to needlessly die?](#)



William Eden @WilliamAEden · 16h

...

This sounds like hyperbole, but it's absolutely not. In what world is a drug so effective it is *unethical to continue the trial* and yet it *is ethical* to take several weeks or more for the regulators to sit down and tell us we can use it? Hint: there is no such world.



Eliezer Yudkowsky ✅ @ESYudkowsky · 16h

FDA plans to kill 50,000 Americans over the next month - over 16X the direct casualties from 2001/9/11 - by keeping it illegal for doctors to prescribe Paxlovid. Can no one stop these homicidal supervillains?
astralcodexten.substack.com/p/when-will-th...

[Show this thread](#)

Will is right. This is not hyperbole. It sure looks like a lot of people are going to needlessly die.

This is a *system of anti-ethics*, the [reverse](#) of actual ethics, because supporting rules in opposition to ethics [shows your loyalties won't be threatened](#) by concerns about actual ethics.²

It is not hyperbole, but it is *imprecise*. What's our best guess of how many Americans will die as a result of the FDA's inaction on Paxlovid?

For a given person, Paxlovid saves their life *if and only if*:

They get Covid-19 and would have died without Paxlovid.

Covid-19 gets detected early enough to use Paxlovid.

They are given Paxlovid.

They live.

We would need to subtract any opportunity costs or side effects from Paxlovid, but for now it seems reasonable to assume the side effects are small when used on Covid-19 positive patients when compared to the benefits, and that the opportunity costs are also small since Paxlovid should combine well with other known-to-be-effective options.

We need to estimate the number of people who are going to die between now and approval in order to do step one. My guess is that 50,000 is somewhat high, since deaths lag and the surge is only starting now, so *if we are starting now* I'd use more like 40,000. But it's better to start *at least two weeks ago* if we're computing the overall death toll, so it's more like 60,000 for those purposes. For now, we're going to assume that 'as soon as the clinical trial data is clear as day plus enough time to call and have a meeting' is a reasonable starting point.

Step four only works about 90% of the time, so we need to multiply the number of dead people by 90%, as the other 10% couldn't have been saved. That's the other easy one.

Then we need to figure out how many of those who die detect Covid-19 fast enough to get Paxlovid when it would be effective. That's trickier, especially with our current lack of tests, although that's *also* the FDA's fault. The failure rate that matters is that among cases that

would be fatal without Paxlovid, so there's biases pointing both ways - cases that are harder to detect are mostly harmless, but cases where the patient doesn't seek treatment fast enough tend to be deadlier. Given that our best guess is that most cases are never detected, the first consideration dominates. Also the existence of Paxlovid would make people more inclined to get tested. What is our best guess as to the Paxlovid-world's detection rate? I'm going to guess this is something like 90% as well, the window is reasonably wide, but I'd also believe substantially lower, so I'll use 80% to cover model uncertainty.

That leaves the big question. If the FDA approved Paxlovid, how many doses would be available? Would they be distributed to the people who needed them? Would those doses represent doses that would not otherwise be used, or would they move consumption forward that would have happened anyway?

This is trickier, and I don't have the answer if we look only using causal decision theory on this particular choice. Paxlovid is 90% effective but the effectiveness *per pill* is many orders of magnitude lower since most people would recover anyway, and there were only 7 deaths in the entire clinical trial control group despite them enrolling over 1,000 patients.

It's entirely possible that, while making Paxlovid widely available today instead of widely available on January 1 would save about 30,000 lives, *given the delays already built into the system*, the delay on approval won't *on net* cost many lives. If it's true that [Pfizer will have about 180,000 treatments by end of year](#) (Tweet was corrected), [and each patient takes ten pills](#), then we could only treat 180,000 patients, and that probably saves about 5,000 lives (depends on how selective you think we can be in finding a vulnerable group, we can do better than the study but earlier my estimate of this was clearly too high and I was thinking 20k), less if the bottleneck would have continued for a while, but given how close it is the bottleneck presumably wouldn't last too much longer. Which is only a moderate decrease.

This is relevant to the value of doing an end run of sorts around this particular decision point, or leaning on the FDA regarding it, if you don't count impacts on future other decisions.

Responding to Incentives

That still treats the manufacturing process as if it couldn't have gone any faster, or at least that the decisions made in how fast to go weren't interlinked with expected future FDA behavior, at least in the past. If Pfizer would have expected to FDA to approve Paxlovid a month sooner, that's a strong reason to get more pills made faster, even when there isn't yet full confidence that the drug is going to work. Similarly, an advance commitment to paying a higher price for the first however many pills delivered quickly, conditional on the drug working, and ideally to also cover some costs if it failed, would have done a similar thing.

It's a magician's trick. The FDA causes the delays, and not in a strange [functional decision theory](#) or [acausal trade](#) kind of way. They do it in the direct obvious kind of way. It's worth quoting the rest of Eliezer's thread.



Eliezer Yudkowsky ✅ @ESYudkowsky · 6h

Congrats to the FDA for producing an equilibrium where producers time their manufacturing capex for after they're sufficiently confident of approval, timed no sooner than they expect approval, and people think this means the FDA is not counterfactually responsible.

2

8

75



Eliezer Yudkowsky ✅ @ESYudkowsky · 6h

I suppose this does make things in fact more complicated than they would be if Pfizer already had a warehouse full of pills they couldn't sell. Good job FDA, no hope of political action now that you've made things more complicated than absolute stark simplicity.

1

2

22



Eliezer Yudkowsky ✅ @ESYudkowsky · 6h

Next up: the impossible job of explaining that "manufacturing difficulties" are not

sheer unalterable exogenous facts, which is why it somehow always ends up taking what seems like a "reasonable amount of time" to a pharma executive given the political environment

2

1

32



Eliezer Yudkowsky ✅ @ESYudkowsky · 6h

If manufacturing times were exogenous, they'd sometimes take 6 days or 6 years instead of 6 months. And the part where they don't just run a prediction market and start 6 months earlier is because of regulated compensation structure. But now it's complicated so they win.



Eliezer Yudkowsky ✅ @ESYudkowsky · 6h

So let's try a simple principle. If someone is going to die and it's illegal to give them a known medicine that will save them, the legal system is committing murder by withholding medicine, regardless of how the market in medicines then doesn't manufacture it faster.

4

5

44



Eliezer Yudkowsky ✅ @ESYudkowsky · 6h

If you don't want to be a murderer, let healers treat patients. Because of the way the market treats pills it might not be able to sell, and how politics treats issues that sound complicated, this "simplistic" moral stance is the only viable one.

1

3

28



Eliezer Yudkowsky ✅ @ESYudkowsky · 6h

Replying to @ESYudkowsky

A known cure exists. It's illegal. 50,000 people will die without it. So they were murdered. Yes, all of them, the left-side head is right about how this works, even though how the system effectuates manufacturing times can only be understood by the hooded head.

If one is counting the [invisible graveyard](#) now that it is about to both grow far larger and be much more highly visible, one needs to *at least* count anyone denied existing known-to-be-safe-and-effective life-saving medicine, and those for whom denial is the result of expected delays and compensation schemes preventing timely mass production. The question is the extent to which one counts those for whom treatments were never developed, or were only developed after they already died, because of the *expectation* of FDA inaction.

Thus, my answers are:

For an unreasonable narrow interpretation that only counts those for whom the medicine was already sitting in a warehouse waiting for approval, and treat that shortage as a 'whoops, making things is hard and takes time' rather than a directly caused effect, the FDA is going to directly murder about 5,000 people in the United States.

For a compromise position where we count those who died while the FDA was holding up approval that would have been saved, the FDA is going to murder about 30,000 people in the United States if you start now, somewhat more if you start when we learned Paxlovid was safe and effective minus reasonable time to confirm this.

For a position that is comparing to the counterfactual of what would have happened if Pfizer knew in March 2020 that any pill it developed would be given out whenever there was sufficient Bayesian evidence it was safe and effective, and that they would be well-compensated for having more pills available quickly, the death toll is much higher, and especially much higher overseas.

As an exercise, one might want to consider the same calculations to delays in approval of the vaccines, remembering that Moderna and Pfizer both had their vaccine candidates within two days of starting work. And one could also do the same calculation on the *continuing* failure to approve and allow tests. Or one could include all the psychological and economic damage that could have been prevented, or the lives not fully lived, while we are forced to wait.

Footnotes (not working right yet, sorry):

1

This is not Scott's language, nor is it to ascribe murderous madness to any particular person. It is still important to call things by their right names.

2

One may wish to consider that perhaps this system of anti-ethics is part of a broader system that is also anti-epistemic, anti-virtue and anti-value in general, and that there is what might be called The Implicit Conspiracy that implicitly cooperates with and rewards with power those who demonstrate their dedication to this reversed system, and punishes those who fail to demonstrate this dedication or reward those who do show it. Would the resulting world look like our world?