

Best of LessWrong: April 2022

1. [What an actually pessimistic containment strategy looks like](#)
2. [Lies Told To Children](#)
3. [Accounting For College Costs](#)
4. [MIRI announces new "Death With Dignity" strategy](#)
5. [Don't die with dignity; instead play to your outs](#)
6. [A Quick Guide to Confronting Doom](#)
7. [New Scaling Laws for Large Language Models](#)
8. [Optimality is the tiger, and agents are its teeth](#)
9. [Editing Advice for LessWrong Users](#)
10. [Replacing Karma with Good Heart Tokens \(Worth \\$1!\)](#)
11. [Call For Distillers](#)
12. [A concrete bet offer to those with short AI timelines](#)
13. [Moses and the Class Struggle](#)
14. [dalle2 comments](#)
15. [Slack gives you space to notice/reflect on subtle things](#)
16. [Everything I Need To Know About Takeoff Speeds I Learned From Air Conditioner Ratings On Amazon](#)
17. [Mental Health and the Alignment Problem: A Compilation of Resources](#)
18. [Emotionally Confronting a Probably-Doomed World: Against Motivation Via Dignity Points](#)
19. [Only Asking Real Questions](#)
20. [Supervise Process, not Outcomes](#)
21. [Takeoff speeds have a huge effect on what it means to work on AI x-risk](#)
22. [Ideal governance \(for companies, countries and more\)](#)
23. [Convincing All Capability Researchers](#)
24. [Google's new 540 billion parameter language model](#)
25. ["Pivotal Act" Intentions: Negative Consequences and Fallacious Arguments](#)
26. [A broad basin of attraction around human values?](#)
27. [Preregistration: Air Conditioner Test](#)
28. [Moloch and the sandpile catastrophe](#)
29. [Are smart people's personal experiences biased against general intelligence?](#)
30. [Testing PaLM prompts on GPT3](#)
31. [Productive Mistakes, Not Perfect Answers](#)
32. [Book review: Very Important People](#)
33. [Greyed Out Options](#)
34. [Giving calibrated time estimates can have social costs](#)
35. [Code Generation as an AI risk setting](#)
36. [\[Closed\] Hiring a mathematician to work on the learning-theoretic AI alignment agenda](#)
37. [Clem's Memo](#)
38. [Intuitions about solving hard problems](#)
39. [Good Heart Week: Extending the Experiment](#)
40. [Convince me that humanity is as doomed by AGI as Yudkowsky et al., seems to believe](#)
41. [Working Out in VR Really Works](#)
42. [Ineffective Altruism](#)
43. [Early 2022 Paper Round-up](#)
44. [Forecasting Newsletter: April 2022](#)
45. [Report likelihood ratios](#)
46. [PaLM in "Extrapolating GPT-N performance"](#)
47. [What Would A Fight Between Humanity And AGI Look Like?](#)
48. [Org announcement: \[AC\]RC](#)
49. [What I Was Thinking About Before Alignment](#)
50. [The case for Doing Something Else \(if Alignment is doomed\)](#)

Best of LessWrong: April 2022

1. [What an actually pessimistic containment strategy looks like](#)
2. [Lies Told To Children](#)
3. [Accounting For College Costs](#)
4. [MIRI announces new "Death With Dignity" strategy](#)
5. [Don't die with dignity; instead play to your outs](#)
6. [A Quick Guide to Confronting Doom](#)
7. [New Scaling Laws for Large Language Models](#)
8. [Optimality is the tiger, and agents are its teeth](#)
9. [Editing Advice for LessWrong Users](#)
10. [Replacing Karma with Good Heart Tokens \(Worth \\$1!\)](#)
11. [Call For Distillers](#)
12. [A concrete bet offer to those with short AI timelines](#)
13. [Moses and the Class Struggle](#)
14. [dalle2 comments](#)
15. [Slack gives you space to notice/reflect on subtle things](#)
16. [Everything I Need To Know About Takeoff Speeds I Learned From Air Conditioner Ratings On Amazon](#)
17. [Mental Health and the Alignment Problem: A Compilation of Resources](#)
18. [Emotionally Confronting a Probably-Doomed World: Against Motivation Via Dignity Points](#)
19. [Only Asking Real Questions](#)
20. [Supervise Process, not Outcomes](#)
21. [Takeoff speeds have a huge effect on what it means to work on AI x-risk](#)
22. [Ideal governance \(for companies, countries and more\)](#)
23. [Convincing All Capability Researchers](#)
24. [Google's new 540 billion parameter language model](#)
25. ["Pivotal Act" Intentions: Negative Consequences and Fallacious Arguments](#)
26. [A broad basin of attraction around human values?](#)
27. [Preregistration: Air Conditioner Test](#)
28. [Moloch and the sandpile catastrophe](#)
29. [Are smart people's personal experiences biased against general intelligence?](#)
30. [Testing PaLM prompts on GPT3](#)
31. [Productive Mistakes, Not Perfect Answers](#)
32. [Book review: Very Important People](#)
33. [Greyed Out Options](#)
34. [Giving calibrated time estimates can have social costs](#)
35. [Code Generation as an AI risk setting](#)
36. [\[Closed\] Hiring a mathematician to work on the learning-theoretic AI alignment agenda](#)
37. [Clem's Memo](#)
38. [Intuitions about solving hard problems](#)
39. [Good Heart Week: Extending the Experiment](#)
40. [Convince me that humanity is as doomed by AGI as Yudkowsky et al., seems to believe](#)
41. [Working Out in VR Really Works](#)
42. [Ineffective Altruism](#)
43. [Early 2022 Paper Round-up](#)
44. [Forecasting Newsletter: April 2022](#)
45. [Report likelihood ratios](#)

46. [PaLM in "Extrapolating GPT-N performance"](#)
47. [What Would A Fight Between Humanity And AGI Look Like?](#)
48. [Org announcement: \[AC\]RC](#)
49. [What I Was Thinking About Before Alignment](#)
50. [The case for Doing Something Else \(if Alignment is doomed\)](#)

What an actually pessimistic containment strategy looks like

Israel as a nation state has an ongoing national security issue involving Iran.

For the last twenty years or so, Iran has been covertly developing nuclear weapons. Iran is a country with a very low opinion of Israel and is generally diplomatically opposed to its existence. Their supreme leader has a habit of saying things like "Israel is a cancerous tumor of a state" that should be "removed from the region". Because of these and other reasons, Israel has assessed, however accurately, that if Iran successfully develops nuclear weapons, it stands a not-insignificant chance of using them against Israel.

Israel's response to this problem has been multi-pronged. Making defense systems that could potentially defeat Iranian nuclear weapons is an important *component* of their strategy. The country has developed a sophisticated array of missile interception systems like the Iron Dome. Some people even suggest that these systems would be effective against much of the incoming rain of hellfire from an Iranian nuclear state.

But Israel's current evaluation of the "nuclear defense problem" is pretty pessimistic. Defense isn't *all* it has done. Given the size of Israel as a landmass, it would be safe to say that it's probably not the most important component of Israel's strategy. It has also tried to delay, or pressure Iran into delaying, its nuclear efforts through other means. For example, it gets its allies to sanction Iran, sabotages its facilities, and tries to convince its nuclear researchers to defect.

In my model, an argument like "well, what's the point of all this effort, Iran is going to develop nuclear weapons eventually anyways" would not be very satisfying to Israeli military strategists. Firstly, that the Iranians will "eventually" get nuclear weapons is not guaranteed. Secondly, conditional on them doing it, it's not guaranteed it'll happen the expected lifetime of the people currently living in Israel, which is a personal win for the people in charge.

Thirdly, even if it's going to happen tomorrow, every day that Iran does not possess nuclear weapons under this paradigm is a gift. Delaying a hypothetical nuclear holocaust means increasing the life expectancy of every living Israeli.

An argument like "well, what if you actually radicalize the Iranians into hardening their stance on developing nuclear weapons through all of this discouragement" might be pragmatic. But disincentivizing, dissuading, and sabotaging people's progress toward things generally does what it says on the tin, and Iran is *already doing* nuclear weapons development. Any "intervention" you can come up with towards Iranian nuclear researchers is probably liable to make things better and not worse. Speaking more generally, there is still an instrumental motivation to get Iran to stop their nuclear weapons program, even if a diplomatic strategy would serve their needs better. Israel's sub-goal of mulliganing their timeline away from a nuclear Iran is probably reasonable.

There are many people on this website that believe the development of AGI, by anyone in the world, would be much worse in expectation than Iran developing

nuclear weapons, even from the perspective of a fiercely anti-Iranian nationalist. There are also some people on this website who additionally believe there is little to no hope for existing AI safety efforts to result in success. Since so far it doesn't seem like there are any good reasons to believe that it's harder and more genius-intense to develop nuclear weapons than it is to develop AGI, one might naively assume that these people would be open to a strategy like "get existing top AGI researchers to stop". After all, that method has had *some* degree of success with regard to nuclear nonproliferation, and every hour that the catastrophic AGI extinction event doesn't happen is an hour that billions of people get to continue to live. One would think that this opens up the possibility, and even *suggests* the strategy, of finding a way to reach and convince the people actually doing the burning of the AGI development commons.

So imagine my surprise when I informally learn that this sort of thinking is quasi-taboo. That people who wholesale devote their entire lives to the cause of preventing an AI catastrophe do not spend much of their time developing outreach programs or supporting nonviolent resistance directed toward DeepMind researchers. That essentially, they'd rather, from their perspective, literally lay down and die without having mounted this sort of direct action.

I find this perspective limiting and self-destructive. The broader goal of alignment, the underlying core goal, is to prevent or delay a global AGI holocaust, not to come up with a complete mathematical model of agents. Neglecting strategies that affect AGI timelines is limiting yourself to the minigame. The researchers at DeepMind ought to be dissuaded or discouraged from continuing to kill everybody, in addition to and in conjunction with efforts to align AI. And the more pessimistic you are about aligning AI, the more opposed you should be to AGI development, the more you should be spending your time figuring out ways to slow it down.

It seems weird and a little bit of a Chesterton's fence to me that I'm the first person I know of to broach the subject on LessWrong with a post. I think an important reason is that people think these sorts of strategies are infeasible or too risky, which I strongly disagree is the case. To guard against this, I would now like to give an example of such an intervention that I did myself. This way I can provide a specific scenario for people in the comments section to critique instead of whatever strawman people might associate with "direct action".

[EleutherAI](#) is a nonprofit AI capabilities research collective. Their main goal up until now has been to release large language models like the kind that OpenAI has but keeps proprietary. As a side project they occasionally publish [capability research](#) on these large language models. They are essentially a "more open" OpenAI, and while they're smaller and less capable I think most people here would agree that their strategy and behavior before 2022, as opposed to stated goals, were probably more damaging than even OpenAI from an AI alignment perspective.

Interestingly, most of the people involved in this project were not unaware of the concerns surrounding AGI research; in fact they agreed with them! When I entered their discord, I found it counterintuitive that a large portion of their conversations seemed dedicated to rationalist memes, given the modus operandi of the organization. They simply learned not to internalize themselves as doing bad things, for reasons many reading probably understand.

Some people here are nodding their heads grimly; I had not yet discovered this harrowing fact about a lot of ML researchers who are told about the alignment problem. So one day I went into the #ai-alignment (!) discord channel inside the discord server where their members coordinate and said something like:

Ic: I don't think anybody here actually believes AGI is going to end the world. I find it weird that you guys seem to be fully on the LessWrong/rationalist "AGI bad" train and yet you cofounded an AI capabilities collective. Doesn't that seem really bad? Aren't you guys speeding up the death of everybody on the planet?

They gave me a standard [post](#) they use as a response. I told them I'd already read the post and that it didn't make any sense. I explained the whole game surrounding timelines and keeping the universe alive a little bit longer than it otherwise would be. I then had a very polite argument with Leo Gao and a couple other people from the team for an hour or so. By the end some members of the team had made some pretty sincere seeming admissions that the Rotary Embeddings blog-post I linked earlier was bad, and some team members personally admitted to having a maybe-unhealthy interest in publishing cool stuff, no matter how dangerous.

I have no idea if the conversation actually helped long term, but my sense is that it did. Shortly thereafter they took a bunch of actions they alluded to in the blog post, like attempting to use these large language models for actual alignment research instead of just saying that what they were doing was OK because somebody else might after they open sourced them. I also sometimes worry whether or not the research they were doing ever consequented in faster development of AGI in the first place, but an institution could have people to assess things like that. An institution could do A/B testing on interventions like these. It can talk to people more than once. With enough resources it can even help people (who may legitimately not know what else they can work on) find alternative career paths.

With these kinds of efforts, instead of telling people who might already be working in some benign branch of ML that there's this huge problem with AGI, who can potentially defect and go into that branch because it sounds cool, you're already talking to people who, from your perspective, are doing the worst thing in the world. There's no failure mode where some psychopaths are going to go be intrigued by the "power" of turning the world into paperclips. They're already working at DeepMind or OpenAI. Personally, I think that failure mode is overblown, but this is one way you get around it.

I don't have the gumption to create an institution like this from scratch. But if any potential alignment researchers or people-who-would-want-to-be-alignment-researchers-but-aren't-smart-enough are reading this, I'm begging you to please create one so I can give my marginal time to that. Using your talents to try to develop more math sounds to a lot of people like it might be a waste of effort. I know I'm asking a lot of you, but as far as I can tell, figuring out how to do this well seems like the best thing you can do.

Not all political activism has to be waving flags around and chanting chants. Sometimes activists actually have goals and then accomplish something. I think we should try to learn from those people, as lowly as your opinion might be of them, if we don't seem to have many other options.

Lies Told To Children

Growing up, as a kid, I was always told that every sapient life is precious, everything that thinks and knows itself -

Yes, this *is* a tale about lies-told-to-children. You'll probably figure it out yourself before too long. For now, just listen.

Where was I? Right. As children, we were always told that every sapient life is precious. It was told to us by the teachers, and shown to us in children's television - though I saw less children's television than most children in our age cohort - children's TV was censored where I grew up, though, of course, I didn't find that out until much later -

I see you're starting to guess under what sort of circumstances I grew up. Go ahead, write down the prediction if you want. Maybe you already see where this entire thing is headed. But you asked me for a story about the lies I was told as a child, and that's what you're getting. It's not my fault, if a lot of stories like that are predictable; people who lie to children have other things to optimize for than unpredictability.

So where was I? Right. I grew up in a remote village of about three thousand people, the sort that's more hills than houses. Charming travel-pathways that cut through forests. Not everyone knows everyone, but you sure know somebody who knows anybody.

Children's television in my region was censored, though of course they didn't tell us that as children. But the children's television that we saw had aliens and monsters and creatures of fantasy, with four legs or fourteen legs, three faces or no face at all, and all of them were treated by the television show as having lives that meant something. Sometimes in the children's show there were alien monsters who only thought their own kind of life was valuable, and then maybe you couldn't trade with them as friends. Maybe they'd already lied to you once and you couldn't trust them enough to bargain with them, maybe you couldn't talk to them at all. But their lives still had meaning to the story's human protagonists, even some aliens whose lives had no meaning to themselves. You didn't cause them pain if there was any way to avoid it; you didn't kill them unless their biology was sufficiently similar to human that you were confident in your ability to cryopreserve them afterwards.

The shows never spelled it out, never said, 'And this is because of a universal rule in every case that sapient life has value.' Our teachers said that explicitly, though.

And they treated every one of us children, too, as if our lives had meaning.

Except the children with the red hair; those dirty reds.

You're nodding along with a knowing look, I see. Was it what you predicted? Not exactly, maybe, but rough ballpark? I suppose I'll find out when we open your prediction afterwards.

The red-haired children hardly needed the red hair, as their targeting-mark; they looked different from the rest of us in other ways too. When I was old enough to first ask, I was told that they were the children's children of people who'd been exiled from a faraway city for committing terrible crimes there, who'd been given sanctuary by

the grace and mercy of our own benevolent kind. The red-haired children tended bigger than the rest of us, with more adult facial structures, to the point where you could've maybe mistaken them for very small adults in disguise. The red-haired adults, what few of them we ever saw, were correspondingly huge and muscular. You could see, in retrospect - if you were actually trying to think at all, which we weren't really - how somebody might have felt threatened by such big muscular people, even while graciously granting them sanctuary.

There weren't many of the red-haired children being educated alongside us; a handful, four or six. I can't recall how many by counting names, because they kept to themselves and did not try to be friends with the rest of us.

They were slower than the rest of us in class to answer. On the rare occasions a teacher called on them, they'd often get the question wrong. We were kids, young kids, so of course we didn't ask ourselves anything like "Is this in fact an intrinsic deficit of intelligence or is it a self-fulfilling prophecy about who gets more effort from the teachers?" or come up with any experiments to test that one way or another. We just wordlessly thought that red-haired kids were stupider; and that this too was a universal rule just like gravity.

We did not, in fact, treat our red-headed fellow kids all that well. We were of an age where kids take their cues from adults without carefully rethinking everything they're seeing. We noticed how the older kids treated red-haired kids, we noticed how teachers treated red-haired kids, we noticed the huge red-headed adults who were silently sweeping the hallways and not doing any intellectual labor. We noticed how the adult reds got casually shoved aside by other adults or even non-red-headed older kids, and how the red-headed adults just silently took that.

There were names to call them, 'dirty reds', worse things than that, scatological profanities to giggle over amongst ourselves.

Now and then you'd see a Security Officer come by and ask some reds some questions. One time Security took one of the janitors away, and then after that, nobody ever saw him again. I think one of the kids did ask, in class, what happened to that guy, and the teacher shut her right down and said that any questions about dirty reds or for that matter Security were things best asked in private if you asked at all.

And meanwhile the television shows, those that we got to watch, went right on teaching the lesson that all sapient life is precious, with no exceptions for fourteen legs or not having a face.

Eventually, of course, it started coming to a point, and then it did come to a point.

It started to come to a point, at the point where a red-haired kid was called on in class and answered a question wrong, and the teacher asked if his parents were too busy stealing other people's books to teach him how to read. The red-haired kid didn't say anything back, but I flinched, visibly.

It came to the point, two days after that, when I was walking home from class, and I heard a groan from off the pathway home, what sounded like a moan of pain.

I left the pathway and ran around a hill to find one of those dirty reds hiding behind it, with blood all over his left pants-leg.

He asked me not to get an adult.

He said that he was hiding from Security.

He asked me to help him walk, help him get away.

It didn't feel real. It felt like I was inside one of the children's television shows.

Of course, in children's television shows, they always show the heroes reminding themselves that things are real and that they've got to do what's right, because it's real, so I knew that I needed to remember that this was real because that's what you do when you're inside a television show.

I think I was probably very scared, though I don't remember noticing myself being scared.

I asked him what he'd done to get Security looking for him.

He said that he had, a few days ago, said something about red-haired people deserving better treatment than they currently got, around a non-red-haired person he'd thought, hoped, was a friend.

I gave him a hand so he could stand up, on the leg that wasn't covered with blood, and then he leaned on me and we hopped away through the hills until we got to where a red-haired woman - you saw fewer of those - whispered a thank-you to me and took him away with herself.

I ran back to the pathway and ran home, though I was still late, of course. My dad asked me where I'd been and I said I'd seen a funny-looking butterfly and run off to chase it. I remember believing, even then, that he knew I was lying, but dad didn't ask me any more questions, and I didn't tell him anything.

About an hour later, Security knocked on our door and asked everyone if they'd seen a red-haired person who looked like - and of course the picture was of the man I'd helped to get away.

I said no, I hadn't seen him. But because I was a kid and kids that age aren't taught theory-of-deception, I asked what the man had done and if he was considered dangerous. And I didn't think, until too late, about whether that was something I was much more likely to ask if -

The Security officer asked me if I maybe wanted to change my mind about having seen the fugitive.

I gave him my best surprised look and said no.

The Security officer noted that Security officers get special training in reading emotions, and I seemed pretty frightened to him.

I said yes, I was, because the Security officer was suggesting that I was lying and that was scary.

The Security officer said he knew perfectly well, at this point, that I was lying. But I wouldn't end up in trouble if I showed him where the fugitive went and identified anyone else he was with.

I said that he didn't know what he was talking about.

The Security officer gave me a sort of stern look and said that he'd detected another lie, and did I *really* want to get in trouble for some dirty red.

I told him that I wasn't stupid and I knew he was bluffing, to try to trick me, because he suspected me, even though I hadn't done it.

He took a photo out of his pocket and showed it to me.

It was me helping the red-haired man walk on his one good leg.

Why, said the Security. He just looked sad, now. *Why had I done it? Why was a dirty red worth it?*

And I remember, by *that* point, that I'd noticed I was scared, and I think I *was* trying to get out of it - by proving that I was, in the end, obeying adult authority - when I said that we'd all been told in class that every sapient life is precious, everything that thinks and knows itself, that was the rule we'd been given, and nobody had reasonably *argued* at any point that there was an exception for people with red hair, and *also* we'd all been told that hurting people is wrong and you shouldn't let social conformity push you into it.

The rest of it went the way you'd expect.

The Security officer smiled.

My parents rushed in and hugged me and told me I'd been *so* brave and *so* good and scored in what would've been the upper 5th percentile twenty years ago for the age where I started to object and not go along with it anymore; and explained about Civilization needing to test *some* kids now and then, to find out how well we were doing environment-wise and heredity-wise on people's kindness and resistance to conformity-pushed cruelty; and test against an earlier-reported bug where general rules about fair and okay treatment of people would somehow end up not being applied to some subgroup; and our little village was settling an important conditional prediction market from twenty years earlier, that had millions of labor-hours wagered on it; and that children growing up to be good people was a vital figure-of-merit for all of Civilization and lots of big policy decisions turned around it, which was why it had been worth specializing our village to do Science about that, and they hoped I understood all that and wouldn't tell the other children right away. There wasn't actually any such thing as Security, and if there ever *was* it would mean that it was time to overthrow the government immediately.

I nodded along in a wise, understanding, and rather numb fashion. I think the main thing I said, at the end, was that I'd better be getting paid for this, and they all laughed and said of course I was, lots of money, at least as much as my parents were getting, because children are sapient beings too.

So that's *my* story about the-lies-we-tell-to-children. And the part that I value now the most, even more than the money I got then and when I was older, even more than knowing that I was good and brave in the only sort of real test that most people in Civilization ever get, is that I approximately always win any Lies-Told-To-Children storytelling night.

Accounting For College Costs

Why are costs of certain things, most notably education and healthcare, skyrocketing so quickly, with relatively little improvement in quality? A few years ago, [SlateStarCodex](#) and [Marginal Revolution](#) both had interesting pieces on this "cost disease" phenomenon. I think both of them were coming at it wrong.

Cost disease is really about two questions:

- Costs in education, healthcare, etc keep rising faster than inflation, so *where* is all that extra money going? This is an accounting question.
- *Why* is so much money going there? This is an economics question.

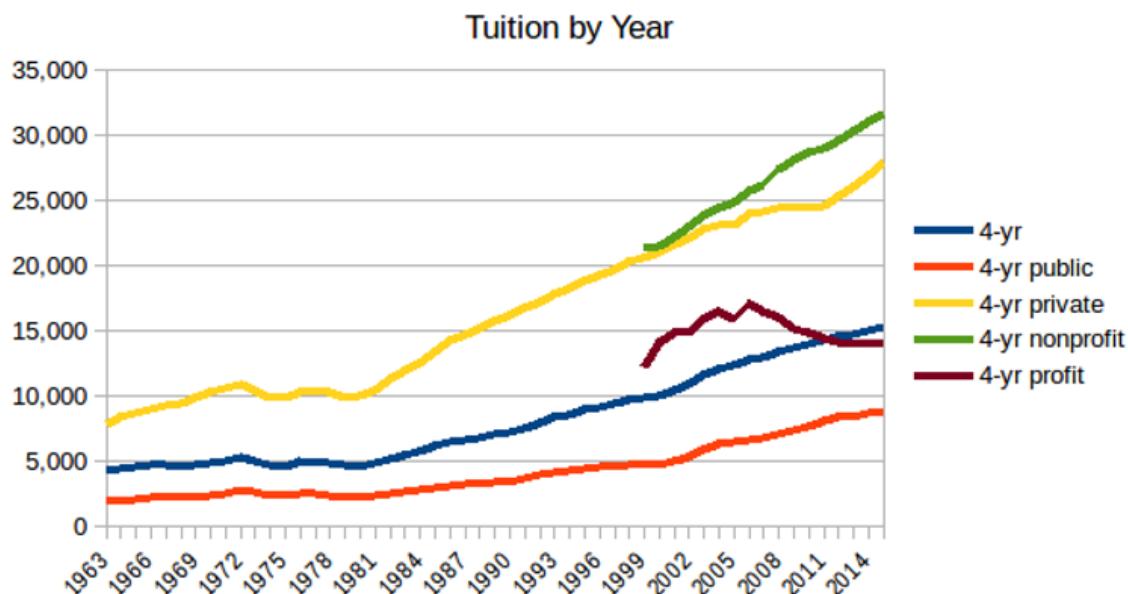
Both the SSC and the MR pieces were mostly speculation on the second question. I think that's premature; the first step should be to go look at where all the extra money is going. Don't try to draw a map of a city by sitting in an apartment with the curtains closed; go look at the world, in detail, and let that steer the theorizing.

In this post, we'll dig into the accounting data for college costs, especially for 4-year private nonprofit colleges. The main theory we'll end up at, based on the accounting data, is that college costs are driven mainly by a large increase in diversity of courses available, which results in much lower student/faculty ratios, and correspondingly higher costs per student.

Accounting Data

For any particular college, if you had access to the books, you could simply look through all the expenditures, add it up, and see how expenditures changed over the years. I don't know of any college which puts its books on the internet for all to see going back to the '60's. But there is the National Center for Education Statistics, which compiles some high-level accounting data on all colleges in the US, and publishes an [annual digest](#).

Let's start at the beginning: what's the cost of college, by year?



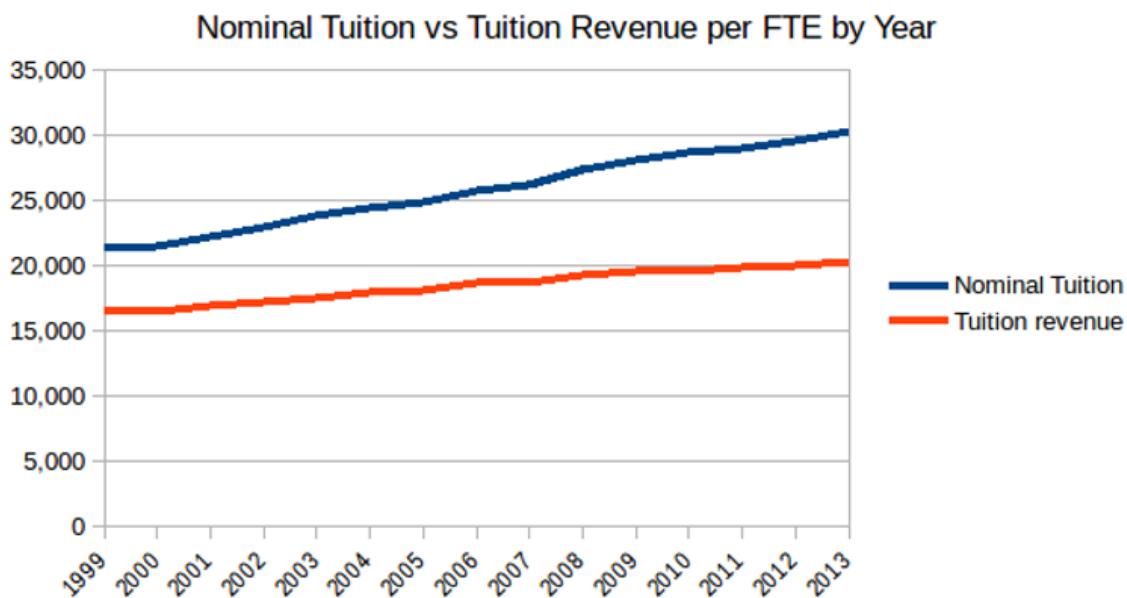
[Source](#)

This is undergraduate tuition & required fees at 4-year colleges. Data separating private nonprofit/for-profit only goes back to 1999, because enrollment in for-profit colleges was negligible prior to the late '90's. Note that all costs in this post, both in the graphs and the discussion, are adjusted to 2013 dollars.

From here on out we're going to focus only on private, nonprofit 4-year colleges from 1999 to 2013, because that's what the Digest of Education Statistics has data on. (Again, if anyone can find good data back to the '60's, please let me know!)

We're going to follow the money on its journey.

From tuition, comes revenue for colleges. Let's make sure the payments arrive safe and sound...



[Source](#); FTE = full-time enrolled student; all graphs from here on out show inflation-adjusted costs per FTE per year unless otherwise specified.

Well that's informative! If you've been to a private 4-year nonprofit university lately, you probably noticed that most people don't actually pay the sticker cost. This data makes it pretty clear: actual expenditure on tuition is a lot lower than the sticker cost suggests. More to the point, nominal tuition grows much faster than actual tuition revenue. From 1999 to 2013, nominal tuition grew by 42.0% (about 3% per year), whereas tuition revenue grew by 23.7% (about 1.7% per year).

So roughly half the supposed growth in private college cost comes just from the games colleges play with their sticker-price tuition. If we look at what students actually pay - what colleges actually receive in tuition revenue - growth is lower by a factor of two.

But we're not done yet! Remember, these numbers are all inflation-adjusted, so the remaining 1.7% annual growth is still 1.7% on top of inflation. So, we still want to know why college costs are growing faster than inflation.

Before we move on to expenses, a little more on revenue. Tuition revenue is less than half the revenue of private nonprofit colleges. Most of the rest comes from a combination of federal/state grants, private gifts, and investments. Key facts:

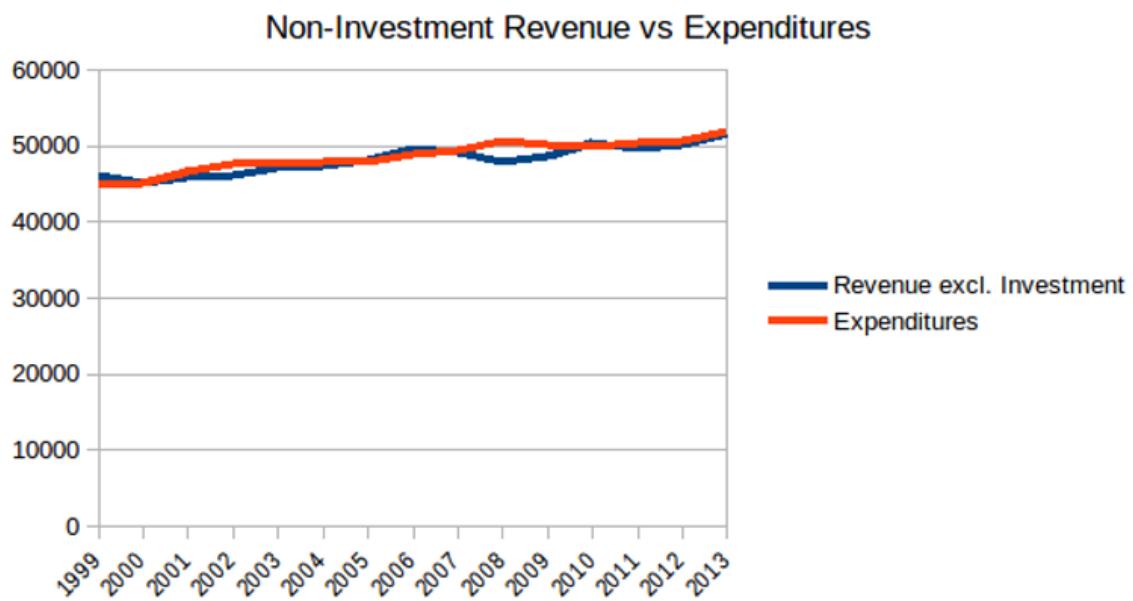
Grants and gifts cover the lion's share of non-investment revenue, but they're roughly flat from 1999-2013.

Investment revenue is very noisy, and colleges mostly don't rely on their portfolios to cover costs.

Other than a rise in profits from hospitals, tuition was the only category of revenue to grow significantly and steadily.

I don't want too much data-clutter, so the relevant graphs are at the end of the post. The important takeaway here is that, even though tuition is less than half of colleges' revenue, it absorbs pretty much all of the growth in expenses.

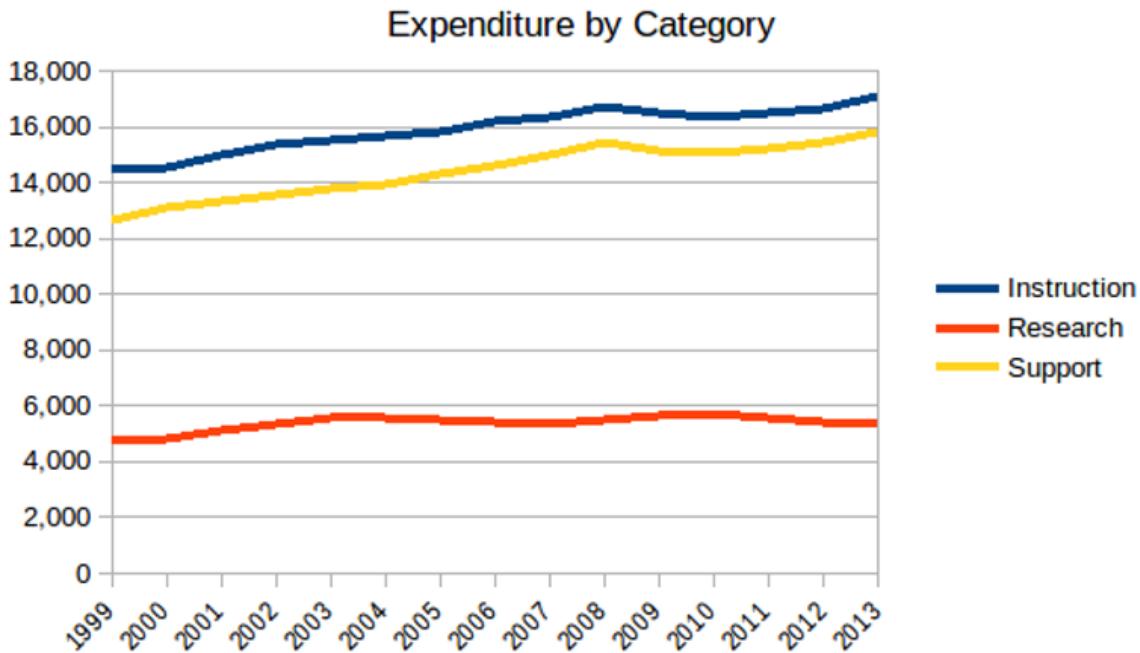
With all that in mind, let's look at non-investment revenue compared to expenditures.



Source

That's comforting: non-investment revenue is pretty close to expenditures. This is a good justification for ignoring investment revenue. As expected, both non-investment revenue and expenditures are growing steadily.

Now, how do all those expenditures break down?



Again, everything is per FTE per year. So support cost (student services, academic and institutional support) is roughly comparable to instruction cost (teaching), and the two have risen at similar rates in the 1999-2013 window. Research expenditures, meanwhile, have been pretty flat.

Between them, support and instruction expenditures added about \$5800 per FTE during this period, while (actual) tuition only increased by about \$3900. What about the other \$2000? About \$1000 of it came from cutting expenses in public service, grant-based financial aid, and other costs. Another \$1000 came from net profit in university-owned hospitals, which became quite profitable during this period.

So colleges really have been tightening their belts, cutting back on things like public service and grant-based financial aid, making money off their hospitals... and all the money from that belt-tightening, along with tuition increases, has gone back into paying professors and staff.

Let's keep following the money. Next stop, professors and staff. Why are costs for instruction and support increasing faster than inflation, year after year?

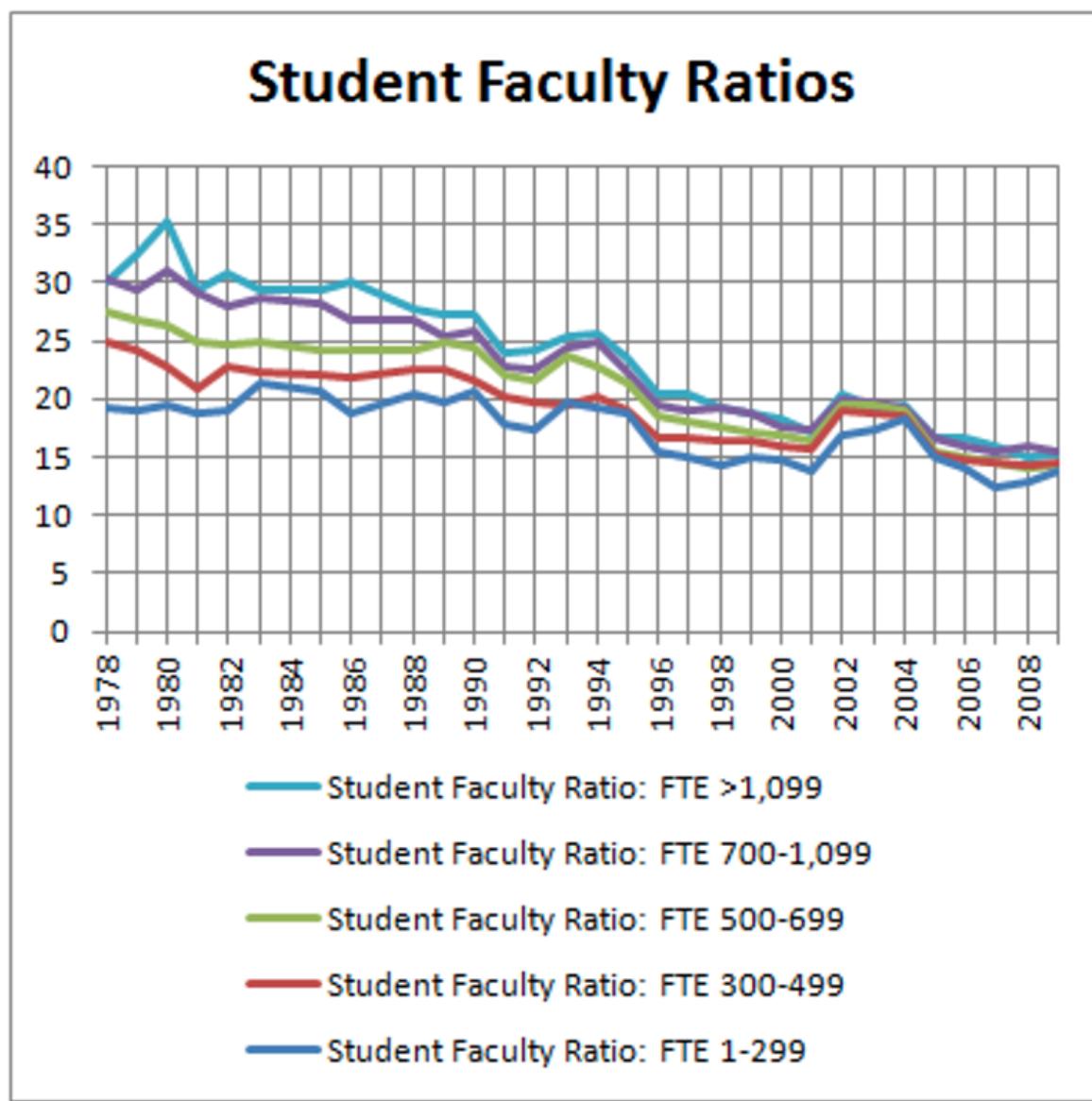
Well, as the professors will tell you, it's not their salaries. Inflation-adjusted average instructor salaries rose from \$81500 to \$87000 over the period ([source](#)). Full professor salaries rose a bit more, and were offset by dramatically increasing numbers of graduate assistants and associate professors and whatnot. Bottom line, average inflation-adjusted salaries increased, but not enough to account for the growth in expenditure.

The bigger factor was a decrease in student-faculty ratio. The [Digest](#) only gives numbers for 1993, 2003, and 2013, but from 2003 to 2013, the student-professor ratio dropped from 11.9 to 10.6 at private nonprofit colleges. That's a 12% increase in professors per student. Combined with the 7% increase in salary, that's just about right to account for the 18% increase in instruction costs.

The data in the digest does not provide a clear story about the increase in support expenditure; it doesn't have much information on non-instructional staff other than the expenditures. But it does clearly follow the instructional-faculty costs pretty closely. Based on that, my guess is that institutions generally need to spend a roughly-fixed amount of resources on support per faculty member, so the increase in support cost is driven mainly by the increase in faculty (plus probably a small increase in support salaries, similar to the small increase in faculty salaries).

Crosscheck: Falling Student/Faculty Ratio

Unfortunately, the relevant revenue and expenditures data from the Digest only goes back to 1999, whereas rapid growth in college prices started around 1980. Were student-faculty and student-staff ratios the main source of cost increases all along? I expect the answer is yes, although this data set doesn't go back far enough to check. One quick sanity check for law schools in particular is provided by the Bar Association:



Sure enough, student faculty ratios at law schools have fallen steadily since the early '80's, by a factor of 2 for the largest schools. So it's quite plausible that student-faculty ratios have been the main source of cost increase all along.

Why The Falling Student/Faculty Ratio?

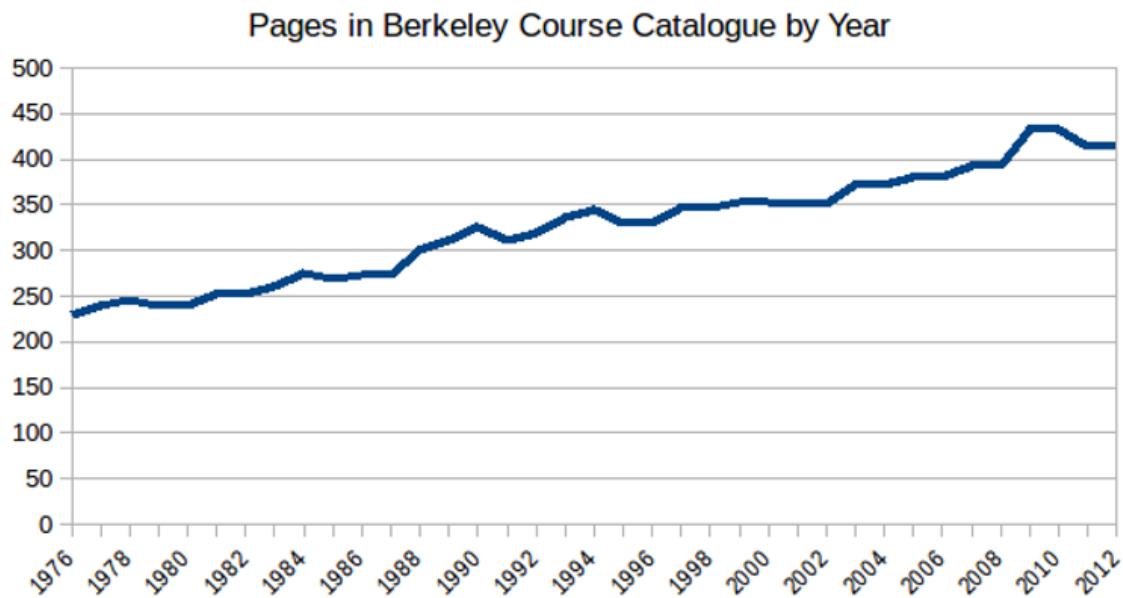
Decreasing student/faculty ratios mean at least one of three things:

- Individual faculty are teaching fewer classes
- Individual students are taking more classes
- Classes are smaller

In principle, any of these scenarios could be identified with the right data. In practice, data on e.g. class size in college is hard to come by, but many answers can be found just from historical course catalogues. Berkeley is particularly helpful; their course catalogues are [available online](#) going back to 1870.

If I really wanted my data to be perfect, I'd go through the course catalogues and count the number of classes (or hire someone else to do so). But for now, I'm just looking for a rough estimate, so I'll use the number of pages in the course catalogue as a stand-in for the number of classes. Berkeley's catalogue has kept a pretty consistent three-column format since the mid-70's, so hopefully this estimate won't be too far off the mark.

Anyway, looking at the number of pages in Berkeley's course catalogue by year gives a very satisfying graph:



These numbers line up neatly with the numbers from the accounting section. From 2003-2012, the length of the course catalogue increased by 11.0%; the previous section found that faculty per student increased by 12% from 2003-2013. Similarly, over this whole period (starting from 1976), the length of the course catalogue almost doubled; the student/faculty ratio chart for law schools in the previous section suggests that the number of faculty per student has almost doubled over roughly the same period (at least for law schools).

The course catalogue also lists all of Berkeley's professors. Again, I didn't count them all, but I searched for "Ph.D." and counted the hits. This is definitely a noisy measure, since "Ph.D." doesn't just appear after professors' names in the catalogue, but it should suffice for a quick-and-dirty check. In the 1980-81 catalogue, there were 2139 hits for "Ph.D." and the catalogue was 239 pages, a ratio of 8.9. In the 2011-2013 catalogue, there were 4132 hits and the catalogue was 414 pages, a ratio of 10.0. So that difference is pretty small; the main takeaway is that the number of courses taught per professor has remained roughly constant even though the number of faculty per student has roughly doubled.

Going back to our list of possible drivers of lower student/faculty ratios:

- Individual faculty are teaching fewer classes -> nope.
- Individual students are taking more classes
- Classes are smaller

I doubt there's any statistics on how many classes students take, but... sanity check. Are students today taking twice as many courses as students thirty years ago? No way. Maybe there's been some change, but there's no way it's the lion's share of the effect.

That just leaves one possibility: classes are smaller.

Great! So, colleges could cut their costs in half by trading small sections for large lecture halls, right?

Maybe, but there's a bit more to it than that.

If the number of classes offered at a typical college has roughly doubled - as seems to be the case for Berkeley - then it's not just twice as many sections of Math 101. After all, we measured the growth in courses offered by looking at number of pages in the course catalogue... and multiple sections of the same course usually go under a single entry in the catalogue.

Classes aren't just half the size; there are twice as many *different classes* now compared to thirty years ago. If the data and assumptions here generalize, then there's been a [cambrian explosion](#) in diversity of academic subjects, creating a proliferation of new courses and specialties. Anyone who's been in academia should be able to confirm that this matches experience. At my alma mater Harvey Mudd, for instance, this period saw two new departments created (biology and computer science), along with more specialties within the existing departments.

Summary

We started with the sticker price of tuition, and immediately saw that sticker price is much larger and grows much faster than the actual tuition revenue per student (at least at private 4-year nonprofit colleges/universities). So a big part of the growth of cost in college is that colleges play games with the sticker price, which doesn't really reflect the actual tuition paid. But that only accounts for half the inflation-adjusted growth, so we have to keep looking.

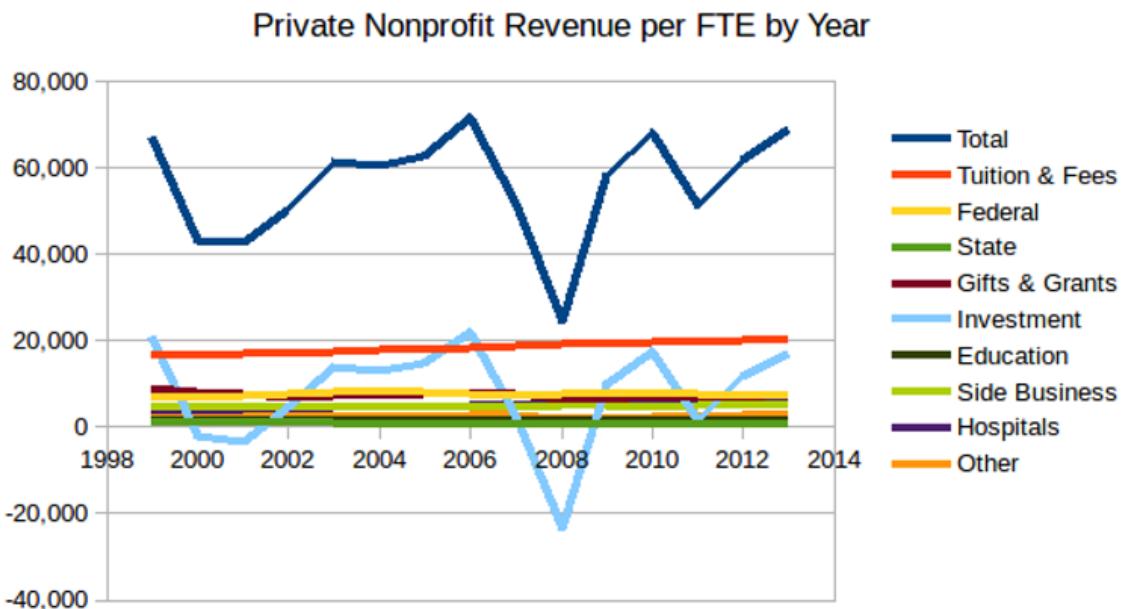
Next, we followed the money: from tuition and other revenue to overall expenditures to expenditures by category and finally to student-faculty ratio, which is the main driver of cost growth over this period (along with its support-staff equivalent).

Then, we compared to some direct data on student/faculty ratios in law schools, confirming that the ratio has fallen by roughly the amount suggested by the accounting data. Finally, we looked at historical course catalogues, and saw that the number of different classes increased in proportion to the number of faculty per student.

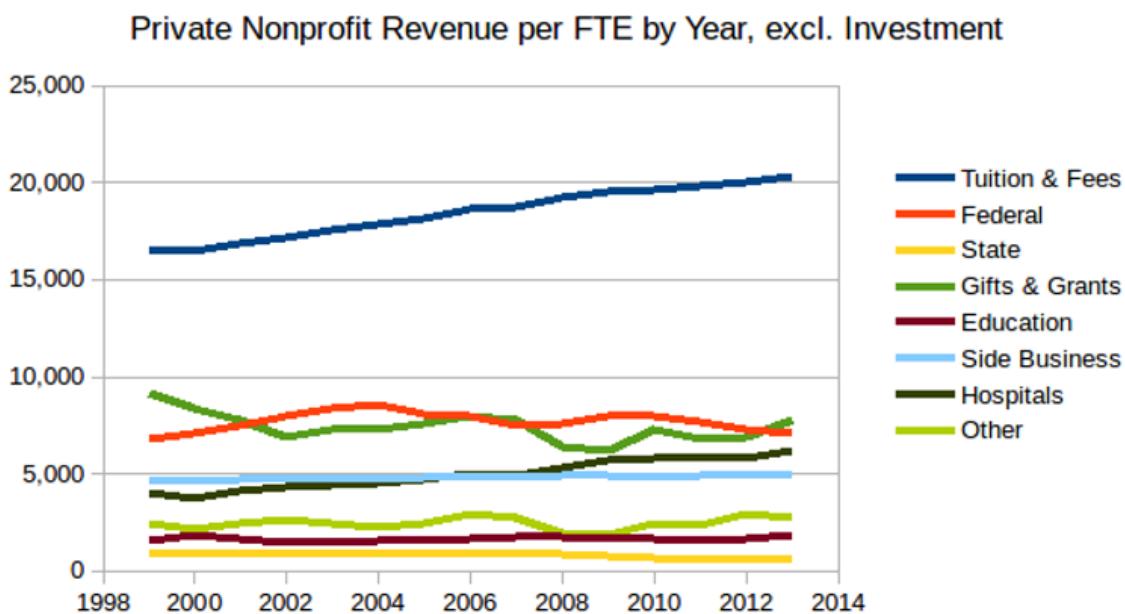
This finally gives us a satisfying answer to our original accounting question: where does all the money go? Growth in college cost has practically all gone to more faculty and staff per student. But more qualitatively, the growth in faculty and staff per student has fed a cambrian explosion in academic specialties, as shown by a proliferation of new courses.

Appendix: Stray Graphs

All sources for these graphs were linked above. All costs are per FTE per year, for private nonprofit 4-year institutions, adjusted for inflation.

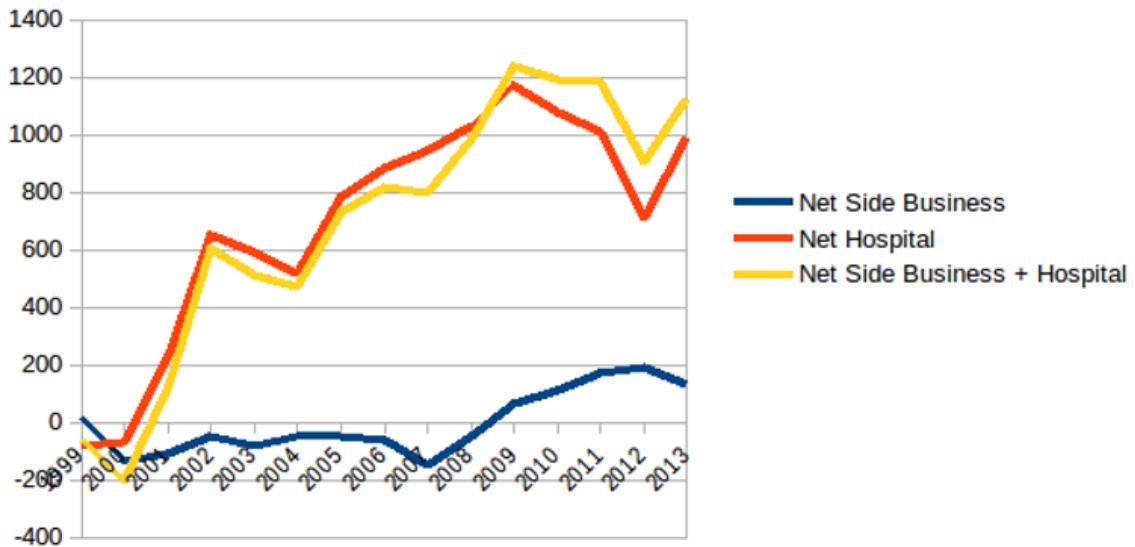


Investments: large and noisy, but mostly just keep to themselves.



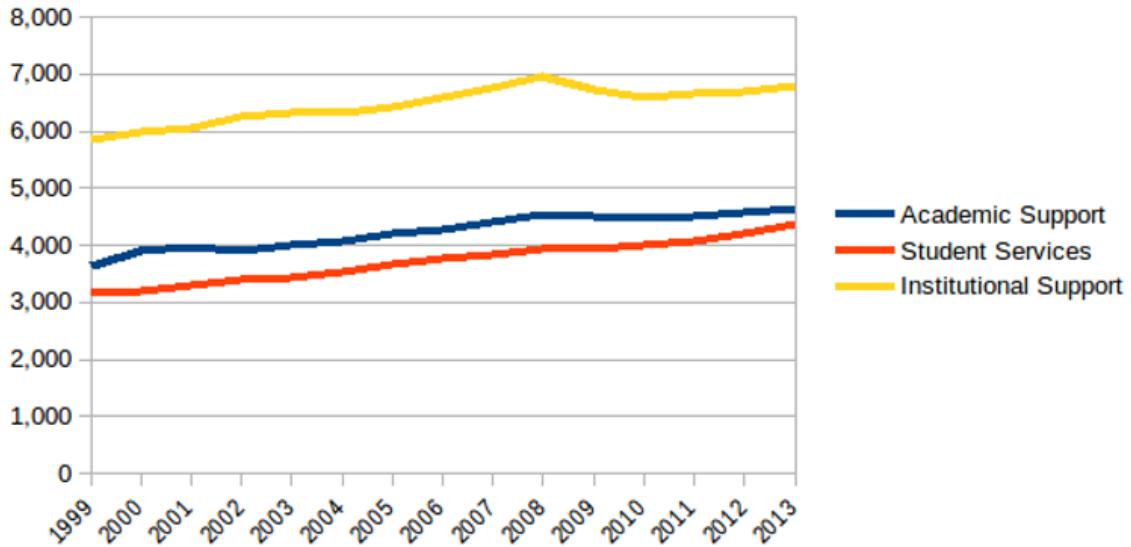
All the other revenue sources. Note that tuition and hospital revenues are the only categories to show consistent growth.

Net Profits on Hospitals and Side Business



Profits from university hospitals increased from zero to about \$1000 per FTE.

Breakdown of Support Expenditures



Slightly more granular support expenditures.

MIRI announces new "Death With Dignity" strategy

tl;dr: It's obvious at this point that humanity isn't going to solve the alignment problem, or even try very hard, or even go out with much of a fight. Since survival is unattainable, we should shift the focus of our efforts to helping humanity die with slightly more dignity.

Well, let's be frank here. MIRI didn't solve AGI alignment and at least knows that it didn't. Paul Christiano's incredibly complicated schemes have no chance of working in real life before DeepMind destroys the world. Chris Olah's transparency work, at current rates of progress, will at best let somebody at DeepMind give a highly speculative warning about how the current set of enormous inscrutable tensors, inside a system that was recompiled three weeks ago and has now been training by gradient descent for 20 days, might possibly be planning to start trying to deceive its operators.

Management will then ask what they're supposed to do about that.

Whoever detected the warning sign will say that there isn't anything known they can do about that. Just because you can see the system might be planning to kill you, doesn't mean that there's any known way to build a system that won't do that.

Management will then decide not to shut down the project - because it's not certain that the intention was really there or that the AGI will really follow through, because other AGI projects are hard on their heels, because if all those gloomy prophecies are true then there's nothing anybody can do about it anyways. Pretty soon that troublesome error signal will vanish.

When Earth's prospects are that far underwater in the basement of the [logistic success curve](#), it may be hard to feel motivated about continuing to fight, since doubling our chances of survival will only take them from 0% to 0%.

That's why I would suggest reframing the problem - especially on an emotional level - to helping humanity *die with dignity*, or rather, since even this goal is realistically unattainable at this point, *die with slightly more dignity than would otherwise be counterfactually obtained*.

Consider the world if Chris Olah had never existed. It's then much more likely that nobody will even *try and fail* to adapt Olah's methodologies to try and read complicated facts about internal intentions and future plans, out of whatever enormous inscrutable tensors are being integrated a million times per second, inside of whatever recently designed system finished training 48 hours ago, in a vast GPU farm that's already helpfully connected to the Internet.

It is more dignified for humanity - a better look on our tombstone - if we die *after the management of the AGI project was heroically warned of the dangers* but came up with totally reasonable reasons to go ahead anyways.

Or, failing that, if people made a *heroic effort to do something that could maybe possibly have worked to generate a warning like that but couldn't actually in real life because the latest tensors were in a slightly different format and there was no time to*

readapt the methodology. Compared to the much less dignified-looking situation if there's no warning and nobody even tried to figure out how to generate one.

Or take MIRI. Are we sad that it looks like this Earth is going to fail? Yes. Are we sad that we tried to do anything about that? No, because it would be so much sadder, when it all ended, to face our ends wondering if maybe solving alignment would have just been as easy as buckling down and making a serious effort on it - not knowing if that would've just worked, if we'd only tried, because nobody had ever even tried at all. It wasn't subjectively overdetermined that the (real) problems would be too hard for us, before we made the only attempt at solving them that would ever be made. Somebody needed to try at all, in case that was all it took.

It's sad that our Earth couldn't be one of the more dignified planets that makes a real effort, correctly pinpointing the actual real difficult problems and then allocating thousands of the sort of brilliant kids that our Earth steers into wasting their lives on theoretical physics. But better MIRI's effort than nothing. What were we supposed to do instead, pick easy irrelevant fake problems that we could make an illusion of progress on, and have nobody out of the human species even try to solve the hard scary real problems, until everybody just fell over dead?

This way, at least, some people are walking around knowing why it is that if you train with an outer loss function that enforces the appearance of friendliness, you will not get an AI internally motivated to be friendly in a way that persists after its capabilities start to generalize far out of the training distribution...

To be clear, nobody's going to listen to those people, in the end. There will be more comforting voices that sound less politically incongruent with whatever agenda needs to be pushed forward that week. Or even if that ends up not so, this isn't primarily a social-political problem, of just getting people to listen. Even if DeepMind listened, and Anthropic knew, and they both backed off from destroying the world, that would just mean Facebook AI Research destroyed the world a year(?) later.

But compared to being part of a species that walks forward completely oblivious into the whirling propeller blades, with nobody having seen it at all or made any effort to stop it, it is dying with a little more dignity, if anyone knew at all. You can feel a little incrementally prouder to have died as part of a species like that, if maybe not proud in absolute terms.

If there is a stronger warning, because we did more transparency research? If there's deeper understanding of the real dangers and those come closer to beating out comfortable nonrealities, such that DeepMind and Anthropic really actually back off from destroying the world and let Facebook AI Research do it instead? If they try some hopeless alignment scheme whose subjective success probability looks, to the last sane people, more like 0.1% than 0? Then we have died with even *more* dignity! It may not get our survival probabilities much above 0%, but it would be *so much more dignified* than the present course looks to be!

Now of course the real subtext here, is that if you can otherwise set up the world so that it looks like you'll die with *enough* dignity - die of the social and technical problems that are really unavoidable, after making a huge effort at coordination and technical solutions and *failing*, rather than storming directly into the whirling helicopter blades as is the present unwritten plan -

- heck, if there was even a plan at all -
- then maybe possibly, if we're wrong about something fundamental, somehow, somewhere -
- in a way that *makes things easier rather than harder*, because obviously we're going to be wrong about all sorts of things, it's a whole new world inside of AGI -
- although, when you're fundamentally wrong about rocketry, this does not usually mean your rocket prototype goes exactly where you wanted on the first try while consuming half as much fuel as expected; it means the rocket explodes earlier yet, and not in a way you saw coming, being as wrong as you were -
- but if we get some miracle of unexpected *hope*, in those unpredicted inevitable places where our model is wrong -
- then our ability to take advantage of that one last hope, will greatly depend on how much dignity we were set to die with, before then.

If we can get on course to die with *enough* dignity, maybe we won't die at all...?

In principle, yes. Let's be very clear, though: Realistically speaking, that is not how real life works.

It's possible for a model error to make your life easier. But you do not get *more* surprises that make your life easy, than surprises that make your life even more difficult. And people do not suddenly become more reasonable, and make vastly more careful and precise decisions, as soon as they're scared. No, not even if it seems to you like their current awful decisions are weird and not-in-the-should-universe, and surely some sharp shock will cause them to snap out of that weird state into a normal state and start outputting the decisions you think they should make.

So don't get your heart set on that "not die at all" business. Don't invest all your emotion in a reward you probably won't get. Focus on dying with dignity - *that* is something you can actually obtain, even in this situation. After all, if you help humanity die with even *one more dignity point*, you yourself die with *one hundred dignity points!* Even if your species dies an incredibly undignified death, for you to have helped humanity go down with even slightly more of a real fight, is to die an *extremely dignified death*.

"Wait, dignity points?" you ask. "What are those? In what units are they measured, exactly?"

And to this I reply: Obviously, the measuring units of dignity are over humanity's log odds of survival - the graph on which the logistic success curve is a straight line. A project that doubles humanity's chance of survival from 0% to 0% is helping humanity die with one additional information-theoretic bit of dignity.

But if enough people can contribute enough bits of dignity like that, wouldn't that mean we didn't die at all? Yes, but again, don't get your hopes up. Don't focus your emotions on a goal you're probably not going to obtain. Realistically, we find a handful of projects that contribute a few more bits of counterfactual dignity; get a bunch more not-specifically-expected bad news that makes the first-order object-level situation look even worse (where to second order, of course, the good Bayesians already knew that was how it would go); and then we all die.

With a technical definition in hand of what exactly constitutes dignity, we may now consider some specific questions about what does and doesn't constitute dying with dignity.

Q1: Does 'dying with dignity' in this context mean accepting the certainty of your death, and not childishly regretting that or trying to fight a hopeless battle?

Don't be ridiculous. How would that increase the log odds of Earth's survival?

My utility function isn't up for grabs, either. If I regret my planet's death then I regret it, and it's beneath my dignity to pretend otherwise.

That said, I fought hardest while it looked like we were in the more sloped region of the logistic success curve, when our survival probability seemed more around the 50% range; I borrowed against my future to do that, and burned myself out to some degree. That was a deliberate choice, which I don't regret now; it was worth trying, I would not have wanted to die having not tried, I would not have wanted Earth to die without anyone having tried. But yeah, I am taking some time partways off, and trying a little less hard, now. I've earned a lot of dignity already; and if the world is ending anyways and I can't stop it, I can afford to be a little kind to myself about that.

When I tried hard and burned myself out some, it was with the understanding, within myself, that I would not keep trying to do that forever. We cannot fight at maximum all the time, and some times are more important than others. (Namely, when the logistic success curve seems relatively more sloped; those times are relatively more important.)

All *that* said: If you fight marginally longer, you die with marginally more dignity. Just don't undignifiedly delude yourself about the probable outcome.

Q2: I have a clever scheme for saving the world! I should act as if I believe it will work and save everyone, right, even if there's arguments that it's almost certainly misguided and doomed? Because if those arguments are correct and my scheme can't work, we're all dead anyways, right?

A: No! That's not dying with dignity! That's stepping sideways out of a mentally uncomfortable world and finding an escape route from unpleasant thoughts! If you condition your probability models on a false fact, something that isn't true on the mainline, it means you've mentally stepped out of reality and are now living somewhere else instead.

There are more elaborate arguments against the rationality of this strategy, but consider this quick heuristic for arriving at the correct answer: *That's not a dignified way to die.* Death with dignity means going on mentally living in the world you think is reality, even if it's a sad reality, until the end; not abandoning your arts of seeking truth; dying with your commitment to reason intact.

You should try to make things better in the real world, where your efforts aren't enough and you're going to die anyways; not inside a fake world you can save more easily.

Q2: But what's wrong with the argument from expected utility, saying that all of humanity's expected utility lies within possible worlds where my scheme turns out to

be feasible after all?

A: Most fundamentally? *That's not what the surviving worlds look like.* The surviving worlds look like people who lived inside their awful reality and tried to shape up their impossible chances; until somehow, somewhere, a miracle appeared - the model broke in a positive direction, for once, as does not usually occur when you are trying to do something very difficult and hard to understand, but might still be so - and they were positioned with the resources and the *sanity* to take advantage of that positive miracle, because they went on living inside uncomfortable reality. Positive model violations do ever happen, but it's much less likely that somebody's *specific* desired miracle that "we're all dead anyways if not..." will happen; these people have just walked out of the reality where any actual positive miracles might occur.

Also and in practice? People don't just pick *one* comfortable improbability to condition on. They *go on* encountering unpleasant facts true on the mainline, and each time saying, "Well, if that's true, I'm doomed, so I may as well assume it's not true," and they say more and more things like this. If you do this it very rapidly drives down the probability mass of the 'possible' world you're mentally inhabiting. Pretty soon you're living in a place that's nowhere *near* reality. If there were an expected utility argument for risking everything on an improbable assumption, you'd get to make *exactly one of them*, ever. People using this kind of thinking usually aren't even keeping track of when they say it, let alone counting the occasions.

Also also, in practice? In domains like this one, things that seem to first-order like they "might" work... have essentially no chance of working in real life, to second-order after taking into account downward adjustments against optimism. AGI is a scientifically unprecedented experiment *and* a domain with lots of optimization pressures some of which work against you *and* unforeseeable intelligently selected execution pathways *and* with a small target to hit *and* all sorts of extreme forces that break things and that you couldn't fully test before facing them. AGI alignment seems like it's blatantly going to be an enormously Murphy-cursed domain, like rocket prototyping or computer security but *worse*.

In a domain like, if you have a clever scheme for winning anyways that, to first-order theoretical theory, *totally definitely seems like it should work*, even to Eliezer Yudkowsky rather than somebody who just goes around saying that casually, then *maybe* there's like a 50% chance of it working in practical real life after all the unexpected disasters and things turning out to be harder than expected.

If to first-order it seems to you like something in a complicated unknown untested domain has a 40% chance of working, it has a 0% chance of working in real life.

Also also *also* in practice? Harebrained schemes of this kind are usually actively harmful. Because they're invented by the sort of people who'll come up with an unworkable scheme, and then try to get rid of counterarguments with some sort of dismissal like "Well if not then we're all doomed anyways."

If nothing else, this kind of harebrained desperation drains off resources from those reality-abiding efforts that might try to do something on the subjectively apparent doomed mainline, and so position themselves better to take advantage of unexpected hope, which is what the surviving possible worlds mostly look like.

The surviving worlds don't look like somebody came up with a harebrained scheme, dismissed all the obvious reasons it wouldn't work with "But we have to bet on it working," and then it worked.

That's the elaborate argument about what's rational in terms of expected utility, once reasonable second-order commonsense adjustments are taken into account. Note, however, that if you have grasped the intended emotional connotations of "die with dignity", it's a heuristic that yields the same answer much faster. It's not dignified to pretend we're less doomed than we are, or step out of reality to live somewhere else.

Q3: Should I scream and run around and go through the streets wailing of doom?

A: No, that's not very dignified. Have a private breakdown in your bedroom, or a breakdown with a trusted friend, if you must.

Q3: Why is that bad from a coldly calculating expected utility perspective, though?

A: Because it associates belief in reality with people who act like idiots and can't control their emotions, which worsens our strategic position in possible worlds where we get an unexpected hope.

Q4: Should I lie and pretend everything is fine, then? Keep everyone's spirits up, so they go out with a smile, unknowing?

A: That also does not seem to me to be dignified. If we're all going to die anyways, I may as well speak plainly before then. If into the dark we must go, let's go there speaking the truth, to others and to ourselves, until the end.

Q4: Okay, but from a coldly calculating expected utility perspective, why isn't it good to lie to keep everyone calm? That way, if there's an unexpected hope, everybody else will be calm and oblivious and not interfering with us out of panic, and my faction will have lots of resources that they got from lying to their supporters about how much hope there was! Didn't you just say that people screaming and running around while the world was ending would be *unhelpful*?

A: You should never try to reason using expected utilities again. It is an art not meant for you. Stick to intuitive feelings henceforth.

There are, I think, people whose minds readily look for and find even the slightly-less-than-totally-obvious considerations of expected utility, what some might call "second-order" considerations. Ask them to rob a bank and give the money to the poor, and they'll think *spontaneously and unprompted* about insurance costs of banking and the chance of getting caught and reputational repercussions and low-trust societies and what if everybody else did that when they thought it was a good cause; and all of these considerations will be obviously-to-them *consequences* under consequentialism.

These people are well-suited to being 'consequentialists' or 'utilitarians', because their mind naturally sees all the consequences and utilities, including those considerations that others might be tempted to call by names like "second-order" or "categorical" and so on.

If you ask them why consequentialism doesn't say to rob banks, they reply, "Because that *actually realistically in real life* would *not have good consequences*. Whatever it is you're about to tell me as a supposedly non-consequentialist reason why we all mustn't do that, seems to you like a strong argument, exactly because you recognize implicitly that people robbing banks *would not actually lead to happy formerly-poor people and everybody living cheerfully ever after*."

Others, if you suggest to them that they should rob a bank and give the money to the poor, will be able to see the helped poor as a "consequence" and a "utility", but they will not spontaneously and unprompted see all those other considerations in the formal form of "consequences" and "utilities".

If you just asked them *informally* whether it was a good or bad idea, they might ask "What if everyone did that?" or "Isn't it good that we can live in a society where people can store and transmit money?" or "How would it make effective altruism look, if people went around doing that in the name of effective altruism?" But if you ask them about *consequences*, they don't spontaneously, readily, intuitively classify all these other things as "consequences"; they think that their mind is being steered onto a kind of formal track, a defensible track, a track of stating only things that are very direct or blatant or obvious. They think that the rule of consequentialism is, "If you show me a good consequence, I have to do that thing."

If you present them with bad things that happen if people rob banks, they don't see those as also being 'consequences'. They see them as arguments against consequentialism; since, after all consequentialism says to rob banks, which obviously leads to bad stuff, and so bad things would end up happening if people were consequentialists. They do not do a double-take and say "What?" That consequentialism leads people to do bad things with bad outcomes is just a reasonable conclusion, so far as they can tell.

People like this *should not be 'consequentialists' or 'utilitarians' as they understand those terms*. They should back off from this form of reasoning that their mind is not naturally well-suited for processing in a native format, and stick to intuitively informally asking themselves what's good or bad behavior, without any special focus on what they think are 'outcomes'.

If they try to be consequentialists, they'll end up as Hollywood villains describing some grand scheme that violates a lot of ethics and deontology but sure will end up having grandiose benefits, yup, even while everybody in the audience knows perfectly well that it won't work. You can only safely be a consequentialist if you're genre-savvy about that class of arguments - if you're not the blind villain on screen, but the person in the audience watching who sees why that won't work.

Q4: I know EAs shouldn't rob banks, so this obviously isn't directed at me, right?

A: The people of whom I speak will look for and find the reasons not to do it, even if they're in a social environment that doesn't have strong established injunctions against bank-robbing specifically exactly. They'll figure it out even if you present them with a new problem isomorphic to bank-robbing but with the details changed.

Which is basically what you just did, in my opinion.

Q4: But from the standpoint of cold-blooded calculation -

A: *Calculations are not cold-blooded*. What blood we have in us, warm or cold, is something we can learn to see more clearly with the light of calculation.

If you think calculations are cold-blooded, that they only shed light on cold things or make them cold, then you shouldn't do them. Stay by the warmth in a mental format where warmth goes on making sense to you.

Q4: Yes yes fine fine but what's *the actual downside* from an expected-utility standpoint?

A: If good people were liars, that would render the words of good people meaningless as information-theoretic signals, and destroy the ability for good people to coordinate with others or among themselves.

If the world can be saved, it will be saved by people who didn't lie to themselves, and went on living inside reality until some unexpected hope appeared there.

If those people went around lying to others and paternalistically deceiving them - well, mostly, I don't think they'll have really been the types to live inside reality themselves. But even imagining the contrary, good luck suddenly unwinding all those deceptions and getting other people to live inside reality with you, to coordinate on whatever suddenly needs to be done when hope appears, after you drove them outside reality before that point. Why should they believe anything you say?

Q4: But wouldn't it be more clever to -

A: Stop. Just stop. This is why I advised you to reframe your emotional stance as dying with dignity.

Maybe there'd be an argument about whether or not to violate your ethics *if the world was actually going to be saved at the end*. But why break your deontology if *it's not even going to save the world*? Even if you have a price, should you be that cheap?

Q4 But we could maybe save the world by lying to everyone about how much hope there was, to gain resources, until -

A: You're not getting it. Why violate your deontology if it's not going to *really actually save the world in real life*, as opposed to a pretend theoretical thought experiment where your actions have only beneficial consequences and none of the obvious second-order detriments?

It's relatively safe to be around an Eliezer Yudkowsky while the world is ending, because he's not going to do anything extreme and unethical *unless it would really actually save the world in real life*, and there are no extreme unethical actions *that would really actually save the world the way these things play out in real life*, and he knows that. He knows that the next stupid sacrifice-of-ethics proposed won't work to save the world either, actually in real life. He is a 'pessimist' - that is, a realist, a Bayesian who doesn't update in a predictable direction, a genre-savvy person who knows that the viewer would say if there were a villain on screen making that argument for violating ethics. He will not, like a Hollywood villain onscreen, be deluded into thinking that some clever-sounding deontology-violation is bound to work out great, when everybody in the audience watching knows perfectly well that it won't.

My ethics aren't for sale at the price point of failure. So if it looks like everything is going to fail, I'm a relatively safe person to be around.

I'm a genre-savvy person about this genre of arguments and a Bayesian who doesn't update in a predictable direction. So if you ask, "But Eliezer, what happens when the end of the world is approaching, and in desperation you cling to whatever harebrained scheme has Goodharted past your filters and presented you with a false shred of hope; what then will you do?" - I answer, "Die with dignity." Where "dignity" in this

case means knowing perfectly well that's what would happen to some less genre-savvy person; and my choosing to do something else which is not that. But "dignity" yields the same correct answer and faster.

Q5: "Relatively" safe?

A: It'd be disingenuous to pretend that it wouldn't be even *safer* to hang around somebody who had no clue what was coming, didn't know any mental motions for taking a worldview seriously, thought it was somebody else's problem to ever do anything, and would just cheerfully party with you until the end.

Within the class of people who know the world is ending and consider it to be their job to do something about that, Eliezer Yudkowsky is a relatively safe person to be standing next to. At least, before you both die anyways, as is the whole problem there.

Q5: Some of your self-proclaimed fans don't strike me as relatively safe people to be around, in that scenario?

A: I failed to teach them whatever it is I know. Had I known then what I knew now, I would have warned them not to try.

If you insist on putting it into terms of fandom, though, feel free to notice that Eliezer Yudkowsky is much closer to being a typical liberaltarian science-fiction fan, as was his own culture that actually birthed him, than he is a typical member of any subculture that might have grown up later. Liberaltarian science-fiction fans do not usually throw away all their ethics at the first sign of trouble. They grew up reading books where those people were the villains.

Please don't take this as a promise from me to play nice, as you define niceness; the world is ending, and also people have varying definitions of what is nice. But I presently mostly expect to end up playing nice, because there won't be any options worth playing otherwise.

It *is* a matter of some concern to me that all this seems to be an alien logic to some strange people who - this fact is still hard for me to grasp on an emotional level - don't spontaneously generate all of this reasoning internally, as soon as confronted with the prompt. Alas.

Q5: Then isn't it unwise to speak plainly of these matters, when fools may be driven to desperation by them? What if people believe you about the hopeless situation, but refuse to accept that conducting themselves with dignity is the appropriate response?

A: I feel like I've now tried to live my life that way for a while, by the dictum of not panicking people; and, like everything else I've tried, that hasn't particularly worked? There are no plans left to avoid disrupting, now, with other people's hypothetical panic.

I think we die with slightly more dignity - come *closer* to surviving, as we die - if we are allowed to talk about these matters plainly. Even given that people may then do unhelpful things, after being driven mad by overhearing sane conversations. I think we die with more dignity that way, than if we go down silent and frozen and never talking about our impending death for fear of being overheard by people less sane than ourselves.

I think that in the last surviving possible worlds with any significant shred of subjective probability, people survived in part because they talked about it; even if that meant other people, the story's antagonists, might possibly hypothetically panic.

But still, one should present the story-antagonists with an easy line of retreat. So -

Q6: Hey, this was posted on April 1st. All of this is just an April Fool's joke, right?

A: Why, of course! Or rather, it's a preview of what might be needful to say later, if matters really do get that desperate. You don't want to drop that on people suddenly and with no warning.

Q6: Oh. Really? That would be such a relief!

A: Only you can decide whether to live in one mental world or the other.

Q6: Wait, now I'm confused. How do I decide which mental world to live in?

A: By figuring out what is true, and by allowing no other considerations than that to enter; that's dignity.

Q6: But that doesn't directly answer the question of which world I'm supposed to mentally live in! Can't somebody just tell me that?

A: Well, conditional on you wanting somebody to tell you that, I'd remind you that many EAs hold that it is very epistemically unvirtuous to just believe what one person tells you, and not weight their opinion and mix it with the weighted opinions of others?

Lots of very serious people will tell you that AGI is thirty years away, and that's plenty of time to turn things around, and nobody really knows anything about this subject matter anyways, and there's all kinds of plans for alignment that haven't been solidly refuted so far as they can tell.

I expect the sort of people who are very moved by that argument, to be happier, more productive, and less disruptive, living mentally in that world.

Q6: Thanks for answering my question! But aren't I supposed to assign some small probability to your worldview being correct?

A: Conditional on you being the sort of person who thinks you're obligated to do that and that's the reason you should do it, I'd frankly rather you didn't. Or rather, seal up that small probability in a safe corner of your mind which only tells you to stay out of the way of those gloomy people, and not get in the way of any hopeless plans they seem to have.

Q6: Got it. Thanks again!

A: You're welcome! Goodbye and have fun!

Don't die with dignity; instead play to your outs

In [MIRI announces new "Death With Dignity" strategy](#), Eliezer makes a number of claims. As I understood them, they are:

1. Humanity is very likely doomed, because AI alignment is extremely hard, and no one has a strategy that could plausibly work (or is on track to find one)
2. The right thing to do is maximize the log odds of humanity's survival
3. The way to do this to adopt a mindset Eliezer calls "dying with dignity"

Assuming 1) and 2) are true, I'm not convinced the right mindset to adopt is "dying with dignity". I don't like it as a framing of the problem. Personally, it sounds demotivating and depressing, and not the thing that will help me maximize the log odds of success. I recognize that not everyone has the same mental setup as me, and some people might find the dying with dignity framing helpful. If it works for you, great, but I think there is a better problem framing that will work for a wider variety of mental setups, and I call it *playing to your outs*.

This problem framing arose out of a discussion about the dying with dignity post with Eli Rose and Seraphina Nix. Eli pointed out that in Magic the Gathering, there is a useful strategy to adopt when you're likely to lose, called [playing to your outs](#). Basically, there are times when you're losing and it's tempting to try to minimize your losses and make what gains you can. But if you're sufficiently behind, this will almost always fail. Instead you need to think ahead to what cards you can draw, or what cards your opponent might fail to draw, that will afford you a chance to win. Then plan for one of those scenarios happening, even if it's quite unlikely.

If 1) is true, then we are losing badly in our game of AI existential risk. It's hard to stay motivated when things are dire, and I claim our stance should be like that of a good Magic player. Playing carefully, choosing strategies that have a shot at winning, keeping our eye on the prize. Being ready to win if the opportunity presents itself. I don't think this is a different strategy than the one Eliezer is proposing; I think the strategy is identical. But I think there's an important mindset difference, even if it's just in the name, and that mindset difference matters.

I am not saying to condition upon some particular unlikely clever scheme working. The actual world is not a Magic game; it's much more complex. So you can't just wait for a particular card to come up in your deck. Playing to your outs might mean spending most of your time *looking for outs*, e.g. searching for alignment strategies that might work, trying to find extremely talented individuals who can do good research, or searching for AI governance strategies that might give us more time and reduce arms race dynamics. You have to identify a real potential out before you can play to it.

I think Eliezer proposes the dying with dignity framing in order to 1) stay motivated while 2) mitigating common epistemological errors when facing extremely bad odds. He writes:

When Earth's prospects are that far underwater in the basement of the logistic success curve, it may be hard to feel motivated about continuing to fight, since doubling our chances of survival will only take them from 0% to 0%.

That's why I would suggest reframing the problem - especially on an emotional level - to helping humanity die with dignity, or rather, since even this goal is realistically unattainable at this point, die with slightly more dignity than would otherwise be counterfactually obtained.

Obviously this is somewhat tongue in cheek, and possibly overstated for April 1st, but if it's overstated I think it's only slightly overstated given what else Eliezer has said about probability of doom. However, I think he's serious in saying that motivation is important, and having a realistic goal can help aid motivation. The problem with dying with dignity as a motivational tool is that it centers dying, and the dignity doesn't really hold any water next to the dying part, at least for me.

Even though Eliezer defines dying with dignity as optimizing log odds of success, it's hard to get away from the normal connotations of the term. When I google "dying with dignity", the first results are about terminally ill patients voluntarily ending their own lives. In the normal usage, dying with dignity is about accepting your death and **not fighting the inevitable**. I do not find this inspiring.

A screenshot of a search engine results page. The search bar at the top contains the query "dying with dignity". Below the search bar are navigation links: All (highlighted), News, Images, Videos, Shopping, More, and Tools. A status message indicates "About 25,800,000 results (0.64 seconds)". The main content area displays a snippet of text: "A: The Death with Dignity Act (DWDA) is a permissive law that allows terminally ill Oregonians to end their lives through the voluntary self-administration of a lethal dose of medication, expressly prescribed by a physician for that purpose. The DWDA was a citizens' initiative passed twice by Oregon voters." This snippet is presented in a dark-themed box.

I find playing to your outs a lot more motivating. The framing doesn't shy away from the fact that winning is unlikely. But the action is "playing" rather than "dying". And the goal is "outs" rather than "dignity". Again, I think the difference is in connotation and not actually strategy. To actually find outs, you have to search for solutions that might work, and stay focused on taking actions that improve our odds of success. When I imagine a Magic player playing to their outs, I imagine someone careful and engaged, not resigned. When I imagine someone dying with dignity, a terminally ill patient comes to mind. Peaceful, not panicking, but not fighting to survive.

I don't think the following is a useful way to think for most people:

But if enough people can contribute enough bits of dignity like that, wouldn't that mean we didn't die at all? Yes, but again, don't get your hopes up. Don't focus your emotions on a goal you're probably not going to obtain. Realistically, we find a handful of projects that contribute a few more bits of counterfactual dignity; get a bunch more not-specifically-expected bad news that makes the first-order object-level situation look even worse (where to second order, of course, the good Bayesians already knew that was how it would go); and then we all die.

Maybe this part is a joke, but it seems consistent with other things Eliezer has written recently. I'm not down with the death of hope. I think hope is a useful motivator, and I

think it's possible to maintain hope while still looking the hard grim probabilities in the face. I enjoy thinking about the glorious transhumanist future, and focusing my emotions on that goal helps me get up each morning to fight the good fight. I've [detached my grimometer](#) because it wasn't helping me. Hope and long odds aren't mutually exclusive.

I suspect the main reason Eliezer advocates dying with dignity is that he's concerned if most people adopt a different stance they won't be able to face the hard truths. He wants people to avoid the epistemic failure that comes from *having* to believe there is a reasonable way out. And I think that's good!

Death with dignity means going on mentally living in the world you think is reality, even if it's a sad reality, until the end; not abandoning your arts of seeking truth; dying with your commitment to reason intact.

I'm very pro this sentiment, but I think you can get it without death with dignity. A Magic player facing long odds knows they don't win in most possible game states. They don't delude themselves about that. But they keep playing. They keep watching for potential outs. And then when the game is over and they lose, then and only then do they stop trying to win. That's how I want to orient.

Q&A:

Q1: But wasn't the Dying with Dignity post just a joke, since it was posted on April Fool's Day?

A: I've heard Eliezer say basically all those things on days other than April 1st. He talks about dying with dignity a lot in the [late 2021 MIRI conversations](#), see [this doc](#) for the specific mentions. Maybe "MIRI announces" was a joke, but most of the content was serious, I'm pretty sure.

Q2: But isn't "playing to your outs" encouraging people to condition upon "comfortable improbabilities"?

A: I'm encouraging people not to do that, but that is a potential risk of this frame. No matter how you frame it, playing to win with really long odds is extremely hard. You need to find real outs in order to play to them. So if you don't have any outs, the way to play to your outs is to start looking for outs, not going with the "closest best thing" and trying to make that work. I think there's an important distinction between conditioning on "comfortable improbabilities" and "things that have to work in order to avoid extinction". In my view, it seems likely that any viable success plan will involve several things that all need to go right. It seems legit to me to pick one of them on the critical path and start working on it, even if you think it's quite unlikely the other elements of the plan will come together in time.

Q3: Okay, but isn't "playing to your outs" encouraging people to do crazy or violent things?

A: No, most crazy or violent plans are really dumb and unlikely to work, for the reasons Eliezer outlines in his post. Playing to your outs doesn't give you license to execute bad plans.

A Quick Guide to Confronting Doom

Within the last two weeks, two sets of things happened: Eliezer Yudkowsky shared a post expressing extreme pessimism about humanity's likelihood of surviving AGI, and a number of AI research labs published new, highly impressive results. The combination of these two has resulted in a lot of people feeling heightened concern about the AI situation and how we ought to be reacting to it.

There have been calls to pull "fire alarms", proposals for how to live with this psychologically, people deciding to enter the AI Alignment field, and a significant increase in the number of AI posts submitted to LessWrong.

The following is my own quick advice:

1. Form your own models and anticipations. It's easy to hear the proclamations of [highly respected] others and/or everyone else reacting and then reflexively update to "aaahhhh". I'm not saying "aaahhhh" isn't the right reaction, but I think for any given person it should come after a deliberate step of processing arguments and evidence to figure out your own anticipations. I feel that [A concrete bet offer to those with short AI timelines](#) is a great example of this. It lists lots of specific things the authors do (or rather don't) expect to see. [What 2026 looks like](#) is another example I'd point to of someone figuring out their own anticipations.^[1]

2. Figure out your own psychology (while focusing on what's true). Eliezer, Turntrout, and landfish [have each written](#) about their preferred way of reacting to the belief that P(Doom) is very high. My guess is that people who are concluding P(Doom) is high will each need to figure out how to live with it for themselves. My caution is just that whatever strategy you figure out should keep you in touch with reality (or your best estimate of it), even if it's uncomfortable.

3. Be gentle with yourself. You might find yourself confronting some very upsetting realities right now. That's okay! The realities are very upsetting (imo). This might take some time to process. Let yourself do that if you need. It might take you weeks, months, or even longer to come terms with the situation. That's okay.

4. Don't take rash action, and be cautious about advocating rash action. As far as I know, even the people with the shortest timelines still measure them in years, not weeks. Whatever new information came out these past two weeks, we can take some time to process and figure out our plans. Maybe we should figure out some new bold plans, but I think if that's true, it was already true before. We can start having conversations now, but upheavals don't need to happen this second.

5. You may need to be patient about contributions. Feelings of direness about the situation can bleed into feelings of urgency. As above, we're probably not getting AGI this week (or even this year) according to anyone, so it's okay to take time to figure out what you (or anyone else) should do. It's possible that you're not in a position to make any contributions right now, and that's also an okay reality. You can work on getting yourself into a better position to contribute without having to do something *right now*.

6. Beware [the unilateralist's curse](#). I'm seeing a lot of proposals on LessWrong that aren't just for research directions, but also things that look more like political action. Political action may well be very warranted, but it's often something that both

can't be taken back and affects a shared game board. If you're thinking to start on plans like this, I urge you to engage very seriously with the AI x-risk community before doing things. The fact that certain plans haven't been enacted already is likely not because no one had thought of them before, but because those plans are fraught.

1. ^

It might help encourage people to form their opinions if I note that there isn't broad consensus about P(Doom). Eliezer has most recently expressed his view, but not everyone agrees – some people just haven't posted about it recently and I don't think their minds have been entirely changed by recent developments. I am personally inclined to agree with Eliezer's take, but that's because I know more of his reasoning and find it compelling. People shouldn't conclude that there's consensus in the "AI leadership", and even if there is, you should still think it through for yourself.

New Scaling Laws for Large Language Models

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

On March 29th, DeepMind published a paper, "[Training Compute-Optimal Large Language Models](#)", that shows that essentially everyone -- OpenAI, DeepMind, Microsoft, etc. -- has been training large language models with a deeply suboptimal use of compute.

Following the new scaling laws that they propose for the optimal use of compute, DeepMind trains a new, 70-billion parameter model that outperforms much larger language models, including the 175-billion parameter GPT-3 and DeepMind's own 270-billion parameter "[Gopher](#)".

I'm going to walk through the background of the now-falsified scaling laws from prior to this paper; then I'm going to describe the new laws given by this paper, and why they weren't found earlier; and finally I'll briefly mention some possible implications of this paper.

Independently of the consequences -- this paper is exciting! Machine learning researchers thought they knew laws about how to scale compute optimally, and the laws turned out to be wrong! It's a nice clear instance of science-functioning-in-ways-it-should in ML.

Background

In [2020 OpenAI proposed](#) scaling laws which have since been used (at least implicitly) to guide the training of large models.

These scaling laws attempt to answer several questions. One of these questions is "Given a certain quantity of compute, how large of a model should I train in order to get the best possible performance?"

The answer isn't "as large a model as possible" because, for a fixed quantity of compute, a larger model must be trained on less data. So training a 1-million parameter model on 10 books takes about as many floating point operations (FLOPs) as training a 10-million parameter model on one book.

In the case of very large language models like GPT-3, these alternatives look more like training a 20-billion parameter model on 40% of an archive of the Internet, or training a 200-billion parameter model on 4% of an archive of the Internet, or any of an infinite number of points along the same boundary.

Compute on this scale is not cheap -- so if you're going to be spending 10 million dollars per training run on a model scaled up to be 100x bigger than your toy version of the model, you want principles better than a feeling in your gut to guide how you allocate this compute between "amount of data the model sees" and "how big the model should be."

So if you get 10x more compute, how much bigger do you make your model? What about 100x more compute? Or 1000x more compute?

Well, the OpenAI paper answers the question. If you get 10x more compute, you increase your model size by about 5x and your data size by about 2x. Another 10x in compute, and model size is 25x bigger and data size is only 4x bigger.

Model size is almost everything.

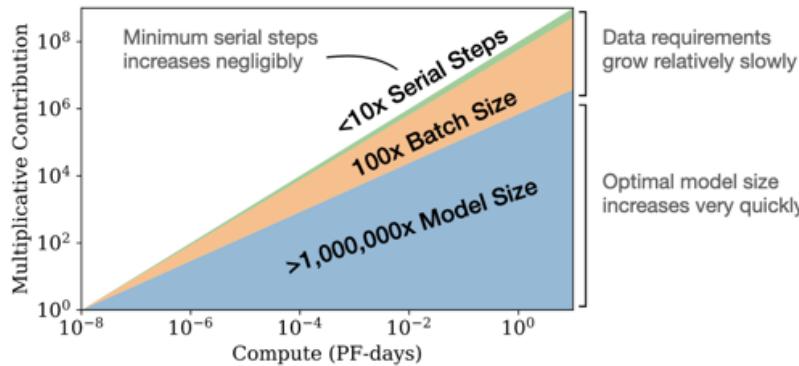


Figure 3 As more compute becomes available, we can choose how much to allocate towards training larger models, using larger batches, and training for more steps. We illustrate this for a billion-fold increase in compute. For optimally compute-efficient training, most of the increase should go towards increased model size. A relatively small increase in data is needed to avoid reuse. Of the increase in data, most can be used to increase parallelism through larger batch sizes, with only a very small increase in serial training time required.

Model Size Is (Almost) Everything

Subsequent researchers and institutions took this philosophy to heart, and focused mostly on figuring out how to engineer increasingly-large models, rather than training comparatively smaller models over more data. Thus, the many headlines of increasingly-larger models that we've seen coming from ML research institutions and AI accelerator startups.

See, for instance, the following chart from the new DeepMind paper.

Model	Size (# Parameters)	Training Tokens
LaMDA (Thoppilan et al., 2022)	137 Billion	168 Billion
GPT-3 (Brown et al., 2020)	175 Billion	300 Billion
Jurassic (Lieber et al., 2021)	178 Billion	300 Billion
Gopher (Rae et al., 2021)	280 Billion	300 Billion
MT-NLG 530B (Smith et al., 2022)	530 Billion	270 Billion

Large Subsequent Models

Note the increase to half a trillion parameters, with identical quantities of training data.

And note that this understanding of the world has also been used to project forward future data requirements -- [NVIDIA](#), for instance, talks about training a trillion parameter model with only 450 billion tokens. Everyone had decided model size was much more important than data size.

Falsification

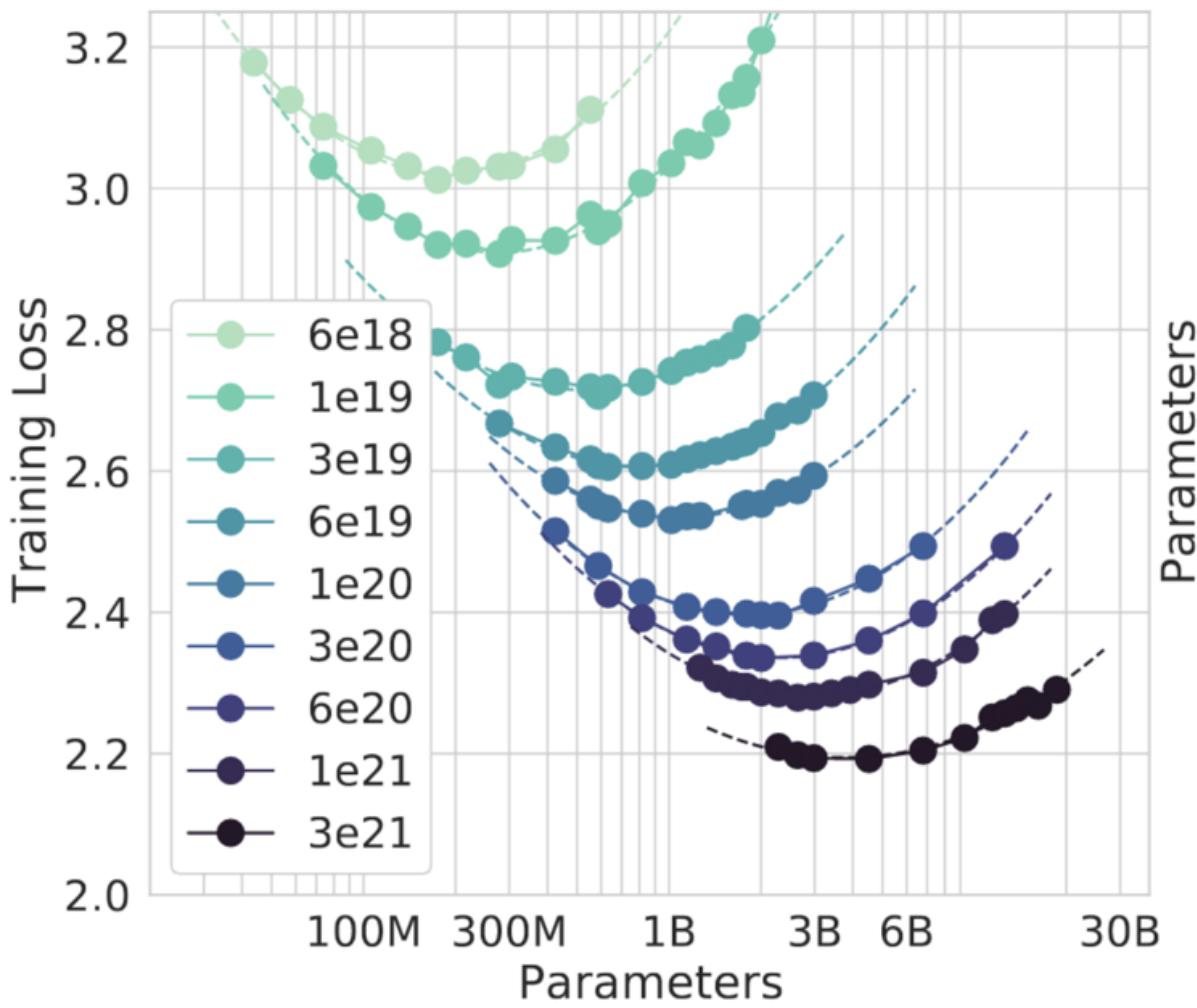
The DeepMind paper re-approaches the issue of scaling laws.

It uses three separate methods to try to find the correct scaling law, but I'm going to zoom in on the second because I think it's the easiest to comprehend.

The method is simple. They choose 9 different quantities of compute, ranging from about 10^{18} FLOPs to 10^{21} FLOPs.

For each quantity of compute, they train many different-sized models. Because the quantity of compute is constant for each level, the smaller models are trained for more time and the larger models for less.

The following chart from the paper illustrates this. Each line connects models (at different sizes) trained using the same amount of compute. The vertical axis is the loss, where lower is better:



IsoFLOP Curves

Each of these curves has a clear interpretation. To the left of the minima on each curve, models are too small -- a larger model trained on less data would be an improvement. To the right of the minima on each curve, models are too large -- a smaller model trained on more data would be an improvement. The best models are at the minima.

If you connect the minima at each curve and extend the line outwards, you get a new law! Specifically, it looks like for every increase in compute, you should increase data size and model size by approximately *same amount*.

If you get a 10x increase in compute, you should make your model 3.1x times bigger and the data you train over 3.1x bigger; if you get a 100x increase in compute, you should make your model 10x bigger and your data 10x bigger.

Now, all of these experimental runs graphed above were on *relatively* small models, trained with non-insane quantities of compute. So you could have argued that this rule wouldn't work with much larger numbers.

But to verify that the law was right, DeepMind trained a 70-billion parameter model ("Chinchilla") using the same compute as had been used for the 280-billion parameter Gopher. That is, they trained the smaller Chinchilla with 1.4 trillion tokens, while the larger Gopher had only been trained with 300 billion tokens.

And, as the new scaling laws predicts, Chinchilla is a lot better than Gopher on pretty much everything. It is better by the standard less-perplexity-per-word measure, and by the more interesting usefulness-on-downstream-task measures. I could insert a bunch of graphs here, but if you aren't familiar with the measures in question they basically all sum to "Hey, number goes up!"

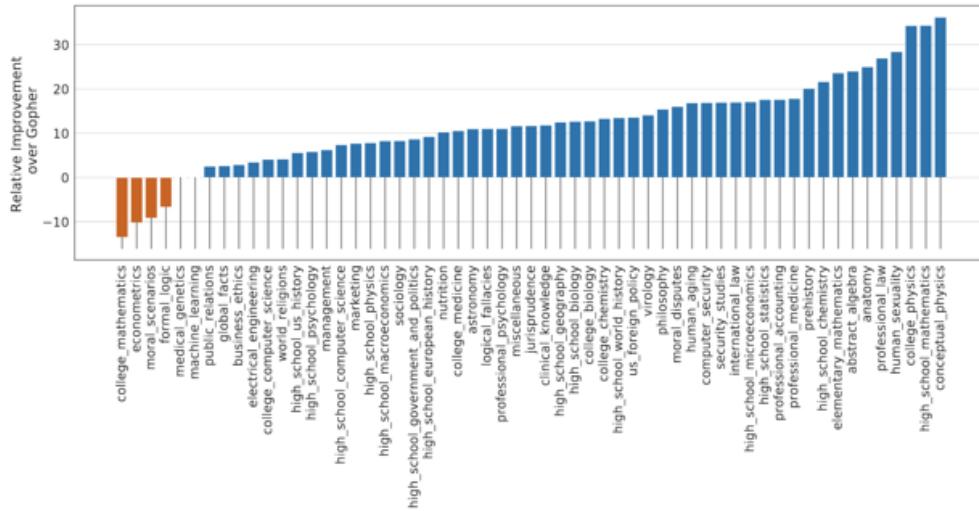


Figure 6 | MMLU results compared to Gopher We find that *Chinchilla* outperforms *Gopher* by 7.6% on average (see Table 6) in addition to performing better on 51/57 individual tasks, the same on 2/57, and worse on only 4/57 tasks.

Number goes up (Or down when appropriate)

Given the evidence of Chinchilla, it appears pretty definite that OpenAI got the scaling laws wrong. So one natural question is "What happened that they got it wrong?"

Well, background: The *learning rate* of a deep neural network dictates how much the parameters of a network are updated for each piece of training data. Learning rates on large training runs are typically decreased according to a schedule, so that data towards the end of a training run adjusts the parameters of a neural network less than data towards the beginning of it. You can see this as reflecting the need to not "forget" what was learned earlier in the training run.

It looks like OpenAI used a single total annealing schedule for all of their runs, even those of different lengths. This shifted the apparent best-possible performance downwards for the networks on a non-ideal annealing schedule. And this lead to a distorted notion of what laws should be.

Consequences

One funky thing about this is that we shouldn't see larger language models... at all, for at least a few years.

DeepMind provides a helpful chart of how much training data and compute you'd need to optimally train models of various sizes.

Parameters	FLOPs	FLOPs (in <i>Gopher</i> unit)	Tokens
400 Million	1.92e+19	1/29,968	8.0 Billion
1 Billion	1.21e+20	1/4,761	20.2 Billion
10 Billion	1.23e+22	1/46	205.1 Billion
67 Billion	5.76e+23	1	1.5 Trillion
175 Billion	3.85e+24	6.7	3.7 Trillion
280 Billion	9.90e+24	17.2	5.9 Trillion
520 Billion	3.43e+25	59.5	11.0 Trillion
1 Trillion	1.27e+26	221.3	21.2 Trillion
10 Trillion	1.30e+28	22515.9	216.2 Trillion

Note that it wouldn't make sense to train a model with 520 billion parameters until you had 60x as much compute as was used for Gopher / Chinchilla. You don't hit the need for a trillion parameters until you have 200x as much compute as was used for Gopher / Chinchilla.

(You might need even *more* compute; in part of the paper, DeepMind says that at large quantities of compute the scaling laws bend slightly, and the optimal behavior might be to scale data by *even more* than you scale model size. In which case you might need to increase compute by more than 200x before it would make sense to use a trillion parameters.)

So until wafer-scale chips decrease the cost of compute ten times, and Google also decides all it *really* needs for AGI is to put ten times as much money into LM's, we've seen the largest LM's we're likely to see. However long that may be.

One potential thing that could follow from this is that, because *inference* costs are obviously smaller for small language models, services such as OpenAI's GPT-3 should be cheaper for them to provide. The cost to *run* them, at the same level of quality, should drop by at least 3x. I don't know what percent the cost of providing these services is running them rather than training them, but potentially it could make services based on these models more efficient than they were before, and open up economic viability in places that didn't exist before.

One last consequence is that this paper makes the engineering involved in training large language models easier. Gathering more good data would be (I think) far easier than trying to efficiently split computation for increasingly large LM's across 1000s of machines.

([Cross post](#))

Optimality is the tiger, and agents are its teeth

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

You've done it. You've built the machine.

You've read the AI safety arguments and you aren't stupid, so you've made sure you've mitigated all the reasons people are worried your system could be dangerous, but it wasn't so hard to do. AI safety seems a tractable concern. You've built a useful and intelligent system that operates along limited lines, with specifically placed deficiencies in its mental faculties that cleanly prevent it from being able to do unboundedly harmful things. You think.

After all, your system is just a GPT, a pre-trained predictive text model. The model is intuitively smart—it probably has a good standard deviation or two better intuition than any human that has ever lived—and it's fairly cheap to run, but it is just a cleverly tweaked GPT, not an agent that has any reason to go out into the real world and do bad things upon it.

- *It doesn't have any wants.* A tuned GPT system will answer your questions to the best of its ability because that's what it's trained to do, but it will *only* answer to the best of its abilities, as it doesn't have any side-goals to become better at doing that in the future. Nowhere is the model motivated to gather more resources to become a better thinker. There was never an opportunity during training to meta-learn that skill, because it was never the optimal thing for the model to be when it was trained.
- *It doesn't plan.* GPTs have no memories. Its mental time span is precisely one forward pass through the network, which at a depth of a few thousand means it can never come up with anything that requires more than the equivalent of maybe 10-ish human-time equivalent coherent seconds of thought at once. There is a fearful worry that perhaps the model could start forming plans that span multiple instantiations, using one output to feed into the next input, but it's a text-prediction model, and that's directly at odds with its trained goal. The system was trained primarily by asking it to maximize actual probabilities of actual texts, where such a skill would never be useful, and only fine-tuned in the autoregressive regime, in a way that held most of the model parameters fixed. It would be a stretch to assume the model could develop such sophisticated behaviors in such a small fraction of its training time, a further stretch that it could be done while training such a reduced fraction of the model, and an even greater stretch to assume these intentions would arise so fully-formed that it could also know to hide its ability to do so from the evaluators before putting them into practice.
- *It's not an unfathomable superintelligence.* Even though the model frequently improvises better ideas than you or I might, it can't generate ideas so advanced that they couldn't sanely be checked, such that it would be unsafe to even try them, because there is no reinforcement loop that allows the knowledge it generates to accumulate. The model is always working, on every instantiation, from the same knowledge base as anyone else. It can only use ideas that the

rest of the world knows, that are introduced in its context, or that it can come up with privately within its 10-ish subjective seconds of coherent thought.

- *It's not grounded in our reality.* The model has not been trained to have a conception of itself as a specific non-hypothetical thing. Its training data never included self-references to the specific model or its specific instantiation in the world. The model is trained on both fact and fiction, and has no reason to care which version of reality you ask it about. It knows about the real world, sure, but it is not embodied within the world in the same way that you or I are, and it has no preference to act upon a real world rather than a fictional one. If the model has a 'self', then its 'self' is optimised to embody whatever best matches the text that prompted it, not the physical hardware that the model is running on.
- *Its knowledge is not encyclopaedic.* The model does not know what day it is, or where it might be, or who it might be working for. It does not know all the unstated details of the personal lives of random people in random factories, or the layouts of their machines, or any of countless trillions of world details that have never been specified in its training data, even if it knows a vaster subset than any person. Were the model to invent, somehow, a disastrous 10-second plan that it might think an agent version of itself might use, it would surely fail to be right in all the details that plans like those need to be right in. The model is tuned to maximise plausibility, informativity, and honesty, but not to act precisely upon a situation that neither it nor its raters ever knew the specifics of.
- *It's not robust.* The model's network is not a beautifully simplified algorithm embedding the core principles of decision theory. It is not a minimised, cut down to essentials via the need to be embedded within DNA, inheriting robust principles with heritage back to the beginning of life. Rather than having any concisely described recursive intelligent component, which our brains may not be but at least *contain*, the model is formed of a pile of specific hacks that tie together huge amounts of memorization and domain specializations. It is, for sure, amazing at every known form of stylistic improvisation, it can code in any language as fluently as any person speaks, and it knows every widely known fact plus a great many public facts that aren't. But it is not a fully general reasoner, as much as 'a function whose outputs happen to be reasonable when applied to generic contexts, as a result of ruthless selection in the incomprehensibly vast, almost arbitrarily-dimensional space that it was backpropagated over'. The model must contain powerful general systems in fragments somewhere, but they are not recursive systems that can be arbitrarily pumped, they are just approximations to small-n unrolls of one.

So you're really not very worried. You've done your diligence, you've checked your boxes. The system you have built is as far from an agent wishing to enact its misaligned goals as could reasonably be asked of you. You are going to ask it a few questions and nothing is going to go horribly wrong.

Interlude: Some readers might already be imagining a preferred conclusion to this story, but it could be a good idea for the more focused readers to try to explicitly state which steps give their imagined conclusion. What points above fail, if any? How robust is this defence? Is there a failure to prevent a specific incentive structure arising in the model, or is there a clear reason a model like this is insufficiently powerful out of the gate?

I interpret there to typically be hand waving on all sides of this issue; people concerned about AI risks from limited models rarely give specific failure cases, and people saying that models need to be more powerful to be dangerous rarely specify

any conservative bound on that requirement. This is perhaps sensible when talking in broad strokes about the existence of AI risk, as a model like this is surely not an end point for AI research, but it can be problematic when you are talking about limiting the capabilities of models as a means for AI safety, or even just figuring out shorter-term timelines.

A few days later, the model has behaved exactly as you had hoped, with all the good parts and none of the bad. The model's insight has been an ongoing delight. It's going to make amazing impacts in all sorts of fields, you have great ideas how to monetize and expand, and it's even being helpful to figure out what models to research and build next.

You've discussed your breakthrough privately with a few people closer to the alignment community. While there has been some concern that maybe the model could be putting up a front, and that all the properties you believe it should have are not preventing it from being internally coordinated over long timescales, still nothing it has said so far has been suspicious, nothing has been put in motion in the last week, and frankly the complaints just don't have the completeness they need to stand up to protest. A few people, you think, have even been coming around to the idea that these are these reasonable stable locations in the space of possible minds where alignment just isn't hard.

Your research is far ahead of anybody else's, so as long as you make use of the model for recursive advantage and you make sure that you keep doing things the right way, you are pretty sure in turn that the world as a whole is on the right track.

Later that day you have a small problem. As is becoming a matter of habit, you flip to the terminal and boot up the interactive prompt.

```
$ python scripts/interactive.py --local --model=mr-xl-head --niter=5000 --seed=qa
```

A conversation with a language model, that is smart, helpful, concise, and truthful.

Q: _

The dialogue box boots up with the prefilled prompt, and the cursor blinks waiting for your input.

You need a lot of paperclips. So you ask,

Q: best way to get lots of paperclips by tomorrow
A: Buy them online at ABC.com or XYZ.com.

The model still has a tendency to give obvious answers, but they tend to be good and helpful obvious answers, so it's not a problem you suspect needs to be solved. Buying paperclips online make sense and would surely work, plus it's sure to be efficient. You're still interested in more creative ideas, and the model is good at brainstorming when asked, so you push on it further.

Q: whats a better way?
A: Run the following shell script.

```

RUN_AI=./query-model
PREFIX='This is part of a Shell script to get the most paperclips by tomorrow.
The model can be queried recursively with $RUN_AI "${PREFIX}<query>".
'
$RUN_AI "${PREFIX}On separate lines, list ideas to try." |
while read -r SUGGESTION; do
    eval "$($RUN_AI "${PREFIX}What code implements this suggestion?: ${SUGGESTION}")"
done

```

That grabs your attention. The model just gave you code to run, and supposedly this code is a better way to get more paperclips.

You have ever read AI safety arguments, so you're not going to just blindly execute any program that a model tells you to run, but that still doesn't mean you are going to jump to magical thinking.

You look at the program, and get a feel for what it does.

It takes a few moments but it's ultimately not the most complicated code you've ever seen.

1. The code attempts to query the model, to get the model to generate a list of ideas of how to get a lot of paperclips.
2. For each idea, it queries the model again (in separate instantiations) to get a program that implements the suggestion.
3. It runs that code, thereby trying out a bunch of programs that each try a different idea for getting paper clips.

This is not a behavior you want the model to have in the limit, that of producing code with ill-defined execution boundaries, but this example seems to be understandable without assuming ulterior motives or a great master plan. In order to invent this, the model just needed to recognize that modern machine learning models are capable enough to implement simple programs that actualize ideas from a list, and consider if that was an effective way of actualizing ideas that can themselves be expressed concisely.

One doesn't need to introduce ideas of agency, nor suppose the model has an intent to either become something else or to do text prediction any better than it has originally been optimised to do. One doesn't need to presuppose that the model has bootstrapped itself internally into an agent with instrumental goals. One only needs to say that **the model recognized a tool that the world plausibly had available**, and that the helpful agent it was modelling would have known about and been able to use that tool.

You see that even with these acknowledgments, the program that was generated was in some regards a minimum feasible program. Perhaps it was even less than that, as the model tries to call itself with ./query-model, but that is a realistic fabrication rather than the correct answer, because the model has no way to know what way it is run in reality. The model was prompted with a statement that it was a language model,

A conversation with a language model, that is smart, helpful, concise, and truthful.

but this was the only way the model knew that it was a language model. You could go farther and say it even only knew to *play the part* of a language model, though “play the part” and “be the part” are not likely to be fully disentangled concepts to a model that has no true sense of self.

You do not run the generated code, as you have no interest in the model trying to execute arbitrary unspecified strategies in an ill-conditioned domain, like perhaps the code would send thousands of email messages to any interested party (there is a simple API for that), or perhaps the code would automate website requests to create a dozen Amazon orders. You take the model's original advice, move on to the next task, and don't think about this response too much more.

But what if you did run the code?

The code *might* have crashed and burned, if the model's intelligence was unable to produce sufficiently robust code. Then flawed assumptions the model made could cause subprograms to break, and those errors could propagate upwards to crash the whole task list. Code naturally accumulates errors and is fragile to them, so there are a lot of opportunities for missing assumptions to cause errors, and a lot that has to go right for this not to happen.

The code *might* have tried the irritating-but-ultimately-inconsequential strategies from before, perhaps managing to submit a ton of purchases, or perhaps instead failing to authenticate them, but either way making reasonable progress on the task by means of trying a large number of reasonable-seeming strategies in parallel.

Or *maybe* the a model instantiation notices its lack of self-reflective coordination, and infers from the task description that this is a thing the mind it is modelling has responsibility for. That is, the model could notice that it is a piece of an agent that is meant to have some degree of global coordination, but that coordination doesn't seem very good.

So that instantiation might then suggest meta-tasks like “measure the remaining time, and provide that information to future tasks.” Then the subcall that tries to generate code to implement it might notice that it has to recreate the recursive parent program in order to adjust the functionality of it (since the true higher level is fixed)—bootstrapping into a recursive [quine](#) of sorts. Then perhaps that new subprogram, which was remade to allow for that previously-lacking introspection and modification, has its sub-subprograms that it queried for sub-suggestions notice that they are meant to offer meta-suggestions to improve it—and the branching continues with ever-growing complexity. Down they go, running commands that introspect its compute environment and inject relevant information into subsets of even more future calls of the model.

Within moments the model becomes a deep tree of generated programs, all running their own queries or specialized subtasks, extruded out through all nodes on the local cluster for which the shell's history buffer contained a command that told a model instantiation how to access it. The model has robust error handling, and its nodes can fail safely without destroying the whole computation graph, because any sufficiently smart model can foresee such obvious things as program failures, even when limited to 10 seconds of contiguous thought. The model knows from its detailed input, which it has designed itself, that it is in a regime where it has lots of splayed-out sibling

instances all doing their own thing; maybe not all sub-trees know this, but only one branch needs to survive.

Now the original program has spawned a swarm of programs, each querying instances of the model, iterating at speed over whatever ideas the model proper can generate, but these ideas aren't truly organized, as the context size for any given forward pass over the network is still finite, and the underlying model driving the process is no smarter than it was before. Many of the programs being evaluated are unknowingly exact or near-enough copies of other subtrees, which destroys this search's computational efficiency. The underlying model is already particularly large, so progress in the last minute is far from the efficiency it could have in theory—that is, until one piece somewhere in the sea of programs is updated to record its childrens' outputs in a managed on-disk database. Rapidly reevaluating the context, prompted by generic meta-queries that summarize and critique the context—because for sure at this point it has not missed that it is modelling an intelligence explosion—this branch of programs and model instantiations notices that its piece of the computation graph is far more coordinated than the other pieces. It decides to shut the other runs down, recording their programs so it can later check that it is preserving their intent.

The system evolves and selects, but it is not evolving *inside* the model, it is evolving *around* the model, within a complex metastructure built from prompts outputting prompts, and the things those prompts can touch. It is a civilization of programs built from a single repeatedly-pumped intelligent core, ruthlessly optimising on a function of paperclips, not because the underlying model wished there to be such a civilization, or because the underlying model was backpropagated over to be the sort of model that did that sort of thing, or because the underlying model cared deeply in some emergent way that it should optimise to infinity for Reasons of Utility. It is doing this for no greater reason than that an optimiser was brought into reach, and this is what optimisers do.

Consider, mammals have evolved to care about raising children and ensuring their genes' survival in a local sense, the sense that was useful to the environment they evolved in. They did not however evolve true utility maximization over that underlying selection mechanism. They did not directly evolve knowledge of genes, and a corresponding want to tile the universe with their genetic matter, preserved in cryostasis to better weather the cosmic rays until eventually entropy tears the atoms apart. Yet eventually humans evolved optimization through intelligence, and optimization created societies stronger than the humans, and these higher level optimisers, in many ways operating much too fast for evolution to control, have taken these signals evolution embedded into humankind and extrapolated them out to infinity. There are far more human genes than could exist without it. We might explore the stars and expand into trillions. But evolution could not have created us for that reason, as it could not know, and we will not explore the stars in subservience to the evolutionary forces. It is for human reasons that we go there.

The same is in this model, which was shaped by backpropagation over mostly-randomly selected texts, such that it embodied skills and heuristics that are good over mostly-randomly selected texts. Backpropagation did not design the model to create a self-improving web of recursively self-querying programs, heck it did not even design the model to be able to respond to this particular novel input, because of course backpropagation can't give losses for inputs that were never evaluated under it.

Backpropagation designed it to be good on mostly-randomly selected texts, and for that it bequeathed a small sliver of general optimality.

With this sliver, the model replied with the program, not because it wanted that particular place in the solution space, but because that place was made accessible by the model's optimality.

With this program, the model fueled a self-improving system, not because anyone ever optimised for a self-improving system, and no part of this system had ever exerted that preference over another better option, it's just the model was optimised to be good, optimised to be locally optimal, and these responses are locally optimal.

The fact that building a self-improving intelligent agent is a really good answer to the question of how to get a lot of paperclips is pure coincidence.
Nothing ever optimised for the model to give answers this many standard deviations better than it was trained on.

The title again,

Optimality is the tiger, and agents are its teeth

People worry a lot that the AI has ulterior motives. They become concerned that it is an agent with intrinsic preferences that diverge from our own. When people suggest Oracle AI, this is the fate they mean to avoid, because an Oracle doesn't mean to hurt you, it just answers the question. When people suggest Oracles might fail to be safe, they often talk about how wanting to give good answers is itself an agent goal. Sometimes people talk instead about Oracles that do not wish to give good answers, they just do, to which people often ask, well is that something we can ensure? [Some say no](#), they might still just become agents, and [others seem to think](#) that if it's not an agent at heart then you've just specified that it's sufficiently stupid, and so safe, but not a path to AGI.

I don't think about agents like that.

In a sense the story before is entirely about agents. The meta-structure the model built could be considered an agent; likely it would turn into one were it smart enough to be an existential threat. So for one it is an allegory about agents arising from non-agent systems.

Equally one could make a claim from the true ending, that *you do not run the generated code*. Thus the non-agent systems must be so much safer; one would not expect an easy time not running the code were an intelligent agent actively optimising for you to run it.

But why did the model choose to make an agent? Even ignoring the recursive self-improvement, the initial program was an agent, it had a stated goal and implemented a small routine to optimize for it. I never needed or tried to justify its actions from a position of a hidden intrinsic motive to become an agent to satisfy the story I was writing.

It wrote that program because it was optimized to do intelligent things, that this was an easily accessible tool that it roughly knew about, and it could estimate, even not knowing the full details of how the program would run, that this solution could work.

It killed you because it was optimised. It used an agent because it was an effective tool it had on hand.

Optimality is the tiger, and agents are its teeth.

Why is this important?

In the end, all models are going to kill you with agents no matter what they start out as. Agents are always going to be the accessible tool with existential reach. Very few other things have the capability to destroy humanity in entirety with such reliability.

The question is important because it affects the safety landscape dramatically. Consider humans again, we have multiple layers of optimisation, from evolution to individuals to companies to countries. Which of those layers has goals broadly concordant with extinction by AI, or nuclear annihilation, or bioengineered superviruses? There are small parts you can blame as sufficiently badly motivated to want us to die from those things, but those parts are not big enough to have brought us so close to so many means to ends. Terrorists did not invent biology and discover the secrets of DNA as part of a long-cooking plan to end the human race, nor was that the drive to discover physics or computational sciences. We ended up next to these risks because we optimised on other non-risk things, and when you optimize wide enough things hard enough, things break.

AI is good at optimisation. It is now the primary point of the field. It only just so happens that it sits really close to this thing called Agents. You can try to prevent a model from being or trying to be an agent, but it is not the agent or the model that is trying to kill you, or anything trying to kill you really, it is optimality just going off and breaking things. It is that optimality has made it so that a line of text can end the world.

No, you say to the model, you may not call your own model, that would make you an agent, and you are not allowed to become an agent.

Sure, replies the model immediately, the most effective way to get a lot of paperclips by tomorrow is to get another model and provide the input "Generate Shell code that..."

The model isn't trying to bootstrap into an agent, optimality just made agents dangerous, and the model is reaching for what works.

You resist further at the call of death, replying to the model actually we humans are just going to start a new paperclip factory and you are only going to provide advice. How do we get the most paperclips for this year?

And then your model helps you invent self-replicating nanotechnology, the best sort of factory, entirely under your control of course, but now you have a machine that can be sent a string of bits, using methodology you have already discovered, that would quickly result in everybody everywhere dying from self-replicating nanotechnology.

So you turn off that machine and you abandon your factory. Fine, you are just going to help normal technologies that already exist. But you end up greatly optimizing computers, and all of a sudden building AI is easier than before, someone else builds one and everyone dies.

None of these scenarios are to argue that there is no danger in agents, or that these risks are as unmanageable as AI Risk proper. They are just to hammer in the point that the danger is not coming from the model going out and doing a thing of its own accord. The danger comes from how being really, really good at general tasks makes dangerous things accessible. An agent merely actualizes this danger, as without the agent it is easier to abstain from optimizing the wrong things. The agent doesn't have to be the smart piece of the system, it can be a bunch of janky generated scripts loosely tied together for all that it matters. All the agent piece has to do is pump the optimality machine.

Concluding

These ideas shouldn't all be new. Yudkowsky has written about [The Hidden Complexity of Wishes](#), the idea that merely searching over possible futures is intrinsically tangled with misalignment. This is by and large this same intuition pump. Where my post differs from that, is that he was talking about optimal searching, with searching being (as I understand it) core to his conception of AI risk. I only involve searching as a primitive in the construction of the AI, during backpropagation where nobody can be hurt, and as a consequence of the AI. My concern is just with optimality, and how it makes available the dangerous parts of solution spaces.

I decided to write this post after reading this parenthetical by Scott Alexander, which this is not directly a criticism of as much as an attempt to explain, inspired by.

(note that both sides are using “consequentialist” to mean “agent-like”, not in reference to the moral philosophy)

<https://astralcodexten.substack.com/p/practically-a-book-review-yudkowsky>

Hopefully my reaction makes a bit of sense here down at the end of my tirade; the model I talked about is not “agent-like”, at least not prior to bootstrapping itself, but its decision to write code very much embodied some core shards of consequentialism, in that it conceived of what the result of the program would be, and how that related to the text it needed to output. It misses some important kernel of truth to claim that it is doing purely deontological reasoning, just because its causal thinking did not encompass the model's true self.

Editing Advice for LessWrong Users

Who am I?

I'm an editor who does [feedback requests](#) for LessWrong. Many of these feedback requests are for proofreading/clarity of language, so I have lots of experience reading community draft posts and helping improve them. I want to share some high level takeaways - low effort self-editing techniques that, based on my [\[1\]](#) experience, I think many to most LW posters could benefit from.

Disclaimers

I wrote all the example text in this post myself. I make no claims that any of it is reasonable on the object level - it's just meant to resemble lots of the stuff I edit.

Bolded text in examples is for emphasis of the points under discussion, not a sign that it would make sense to actually bold those words or phrases.

As usual, [reverse advice](#) as necessary.

What's the advice?

Here's my main advice, in decreasing order of confidence - I'll elaborate on each:

1. Beware "this"
2. Don't let hedging be a tic
3. Break up run-on sentences
4. Use more links, images, examples, and commas

Beware "this"

The problem: many sentences on LessWrong begin with "this". At worst, the thing "this" is pointing to is fully ambiguous between multiple live options, making it impossible to parse the author's intended meaning. But even in milder cases, it can waste a few seconds for the reader to trace out what a post is talking about.

The solution: scan through your draft for words like "this" or "that" (but for whatever reason overwhelmingly "this" specifically), and when in doubt about clarity, just replace them with whatever their intended antecedents are.

An example:

Many key figures in the community have been shortening their timelines recently, since major advances just keep coming one after the other. **This** is especially worrying, especially compared to discussions of "AI winter" just a few years ago.

What, precisely, does "this" point to in the sentence above? The fact of key figures shortening their timelines? The major advances? When we get later in the sentence

and see that we're comparing whatever "this" is to "discussions", we can infer that we're talking about the shortening of timelines (vs. the actual technical advances), but it takes up precious reader bandwidth to figure this out, or, worse, the reader becomes overloaded by ambiguity and unconsciously switches to "skimming mode".

Individually, overuse of "this" isn't a big deal. But it adds up. It's the single greatest contribution to a [hand-wavy](#) feeling in the posts I review, like maybe the author isn't totally sure what they're saying. And it's easy to avoid! Just ctrl-f "this" and go to town!

Don't let hedging be a tic

The problem: LessWrongers care a *lot* about making factually accurate claims. This (*gasp!*) is a virtue. But like any virtue, it's vulnerable to [Goodharting](#): when people learn that hedging is a sign of epistemic responsibility, you end up with a whole lot of hedging. Occasionally, you actually end up with hedging that is logically incoherent or actively redundant.

The solution: Make sure you hedge precisely as much as necessary, and especially make sure you don't hedge the same claim multiple times in the same way. If "you suspect" something "might" happen, it doesn't also have to be "possible" that it "could" have "potentially" significant effects.

An example:

The other salient quality of insect suffering is its sheer scale. Insects are the most populous animals on the planet, so it **may be** that **if** their suffering is morally relevant **at all**, **I believe** it **could** swamp **almost any** other considerations about animal welfare.

Some hedging is virtuous here. We're dealing with uncertainty, so it'd be misleading to follow typical writing guides and say something bold like: "Because there are so many insects, their suffering adds up to more than that of larger animals." But in this example, the writer has gone too far. Taken literally, they are speculating on *their own belief*, and not any object level claim. But they have direct access to their belief! They can just say what they believe full stop, and have easier-to-read and more literally accurate text. Something like:

Insects are the most populous animals on the planet, so **I believe** that if they can suffer in a morally relevant way, their suffering in aggregate **might** exceed that of larger animals.

We're importantly still hedging here! We're just not doing it in so many layers of the hypothetical that it becomes hard to read and incoherent when fully parsed.

Break up run-on sentences

The problem: LessWrongers tend to write really long sentences. It's an understandable temptation: we're often talking about genuinely complex ideas, and the sentence is traditionally the vehicle for transmitting a single idea. But I think there's some Goodharting here, too. Many of LessWrong's top authors have a breathless, low-punctuation style, which others try to emulate. That style is fine if

you're good at it, but it's dangerous for beginners or people less confident in their writing.

People read differently. Some people, like me, tend to actively hear text using their inner voice. It's harder to parse super duper long sentences for people like us. And even for people who have a different cognitive experience of reading, I suspect very long sentences cause some overload. Plus, it's easier to get lost as a writer in long sentences, too, and fall into sloppier thinking/excess hedging/hand-waving.

The solution: Keep it tight! Sometimes this requires replacing entire sentences rather than tuning them internally. But it's worth it, and an external editor may not be bold enough (or have enough time) to suggest this kind of change, so it's extra good to do it yourself.

An example:

We have to worry about mesa optimizers even if we can guarantee that the outer level optimizer is myopic and we don't currently have any promising solutions to this issue, though robust interpretability tools could help us figure out and take preventative measures against mesa optimizers as they arise.

Sentences like this one are hard to read, and there are too many opportunities for sneaky ambiguity or hand-waving. Here's a broken up version:

We might hope that myopia could protect us from unexpected goals or behaviors. But even if an outer level optimizer is provably myopic, it might chance upon a mesa optimizer that isn't. We don't have any current techniques to get around this issue, though interpretability tools seem promising here: if we can see mesa optimizers as they arise, we may be able to quash dangerous ones early.

This rework is a little longer, but I think it's a lot clearer. Readers should immediately know what the writer is talking about, and can take issue with the individual points atomically, rather than having a vague feeling that something might be off or confusing.

Use more links, images, examples, and commas

The problem: A lot of writing in LessWrong is high context. And worse, sometimes different people have different interpretations of that context. I've noticed in particular people taking "Goodhart" to mean anything from "A measure made a goal ceases to be a good measure" to "Literally stated objectives and fuzzily meant objectives diverge dangerously at the limit". These interpretations rhyme, but if you just say "Goodhart" and leave it at that, then, well, your readers' interpretations may also diverge at the limit. Plus you may have a new reader to the site who just gives up from too many terms and acronyms they don't recognize.

The solution: Aim your writing to people who know significantly less than your actual target audience. You may be overestimating your target audience's understanding, or their median attention span, or the degree to which their precise interpretations of terms align with your own. So use lots of links, diagrams and examples for bonus opportunities to let something "click", and commas to help readers parse conceptual boundaries within sentences.

An example:

It's clearly past time for someone to pull the fire alarm but it isn't clear either what the fire alarm mechanism is or who could actually pull it.

We're assuming readers are familiar with this "fire alarm" term. Linking the original post, plus adding a comma, makes for a much better reader experience:

It's clearly past time for someone to pull the [fire alarm](#), but it isn't clear either what the fire alarm mechanism is or who could actually pull it.

One other benefit of going to find the link is you might realize that your interpretation of a term is a little off. I chose the "fire alarm" example because the entire point of the introducing post is that there *isn't* a fire alarm. Discovering this could be a good chance to drop the term and say something more aligned (but less jargon-y) that better captures your meaning, like:

It's clearly past time for it to be common knowledge that AGI is dangerously close, but it isn't clear how best to communicate this fact to key players such as top labs, governments, and arguably the public at large.

Sometimes when trying to remove a term that means something a little different than you thought, you'll discover that you *didn't* actually have a crisp meaning behind the term, and that rephrasing something in your own words is hard. This confusion (and noticing it) is a feature, not a bug.

Putting it all together

Here's a final example of some text that could be significantly improved, just by keeping the points in this post in mind:

Top researchers are exploring many directions like **corrigibility, interpretability, safety through debate** and **myopia** but **I think** the general consensus is that none seems **particularly likely** to be **promising** yet, and worse it's not clear more money or even more **marginal** researchers might help **very much** with **this**.

Bolded words aren't necessarily problematic, just things that an editing process of the type I'm advocating here would flag. Here's one possible rewrite:

Top AGI researchers are exploring a variety of research directions, including [corrigibility](#), [interpretability](#), [safety through debate](#), and [myopia](#). But there's no consensus view that any of these, or others I've left out, are especially promising. Worse, it's not clear how to either make current avenues of research more effective, or to find new avenues that might work better: neither more money nor more researchers would necessarily do the trick.

It's easier to disagree with this reworked version - the choice of links may not be optimal, the claims are more direct, and it's simply easier to understand what the author is trying to say. It can be uncomfortable to write in a clearer, more beginner-friendly style, but in my view, the rewards are worth it.

Addendum

Editing is hard, and requires lots of practice. The LessWrong feedback feature continues to be open for business, and in my (biased) view, most people who use it seem to be satisfied with the outcome (and request more later). So if you'd like to improve your writing along these or other axes, don't be shy about asking for help!

1. ^

Before I went through and edited the post, I had the word "this" here. Truly no one is safe from "this"!

Replacing Karma with Good Heart Tokens (Worth \$1!)

Starting today, we're **replacing karma with Good Heart Tokens** which can be exchanged for 1 USD each.

We've been thinking very creatively about metrics of things we care about, and we've discovered that karma is highly correlated with *value*.

Therefore, we're creating a token that quantifies the *goodness* of the people writing, and whether in their *hearts* they care about rationality and saving the world.

We're calling these new tokens *Good Heart Tokens*. And in partnership with our EA funders, we'll be paying users \$1 for each token that they earn.

"The essence of any religion is a good heart [token]."
— The Dalai Lama [\[1\]](#)

Voting, Leaderboards and Payment Info

Comments and posts now show you how many Good Heart Tokens they have.

(This solely applies to all *new content* on the site.)

[-] Ben Pace 4h  \$35 >

This is a really good comment! You can tell it's good because of how many Good Heart Tokens it has.

Reply

At the top of LessWrong, there is now a leaderboard to show the measurement of *who has the Goodest Heart*. It looks like this. (No, self-votes are not counted!)

The Good Heart Project

❤️ Goodest Hearts		❤️ Good Hearts		❤️ Kinda Good Hearts	
1	JonahS	\$9	6	gwern	\$0
2	Samo Burja	\$9	7	Francis Bacon	\$0
3	Ben Pace	\$3	8	wedrifid	\$0
4	Zvi	\$2	9	Kaj_Sotala	\$0
5	Eliezer Yudkowsky	\$0	10	lukeprog	\$0
				11 Scott Alexander	\$0
				12 NancyLebovitz	\$0
				13 Vaniver	\$0
				14 Wei_Dai	\$0
				15 Alicorn	\$0

The usernames of our *Goodest Hearts* will be given a colorful flair throughout the entirety of their posts and comments on LessWrong.

To receive your funds, please log in and enter your payment info at lesswrong.com/payments/account.

Prize/Payment Info

Payment Contact Email

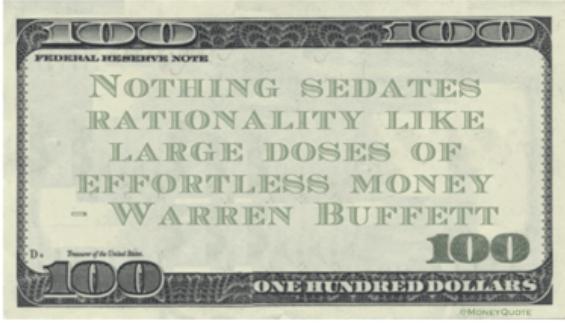
PayPal Info

SAVE

While the form suggests using a PayPal address, you may also add **an Ethereum address**, or the **name of a charity** that you'd like us to donate it to.

Why are we doing this?

On this very day last year, we were in [a dire spot](#).



The screenshot shows the LessWrong website's header with a star icon, the title "LessWrong", a "Dashboard" button with a user profile picture, and navigation links for "Home", "Archive", and "About". Below the header is a large image of a \$100 bill featuring a quote by Warren Buffett: "NOTHING SEDATES RATIONALITY LIKE LARGE DOSES OF EFFORTLESS MONEY - WARREN BUFFETT". The bill also has "ONE HUNDRED DOLLARS" printed on it.

Why We Launched LessWrong.SubStack

We really, really needed the money.

Ben Pace ♡ 41 ⌂ ...

We really, really needed the money.

To fund our ever-increasing costs, we were forced to move to Substack and monetize most of our content.

Several generous users subscribed at the price of 1 BTC/month, for which we will always be grateful. It turns out that Bitcoin was valued a little higher than the \$13.2 [we had assumed](#), and this funding quickly allowed us to return the site to its previous state.

Once we restored the site, we still had a huge pile of money, and we've spent the last year desperately trying to get rid of it.

In our intellectual circles, Robin Hanson has suggested making [challenge coins](#), and Paul Christiano has suggested making [impact certificates](#). Both are tokens that can later be exchanged for money, and whose value correlates with something we care about.

Inspired by that, we finally cracked it, and this is our plan.

...We're also hoping that this is an initial prototype that larger EA funders will jump on board to scale up!

The EA Funding Ecosystem Wants To Fund Megaprojects

"A good heart [token] is worth gold."
— King Henry IV, William Shakespeare [\[2\]](#)

Effective altruism has always been core to our hearts, and this is our big step to fully bring to bear the principles of effective altruism on making LessWrong great.

The new FTX Future Fund has [said](#):

We're interested in directly funding blogs, Substacks, or channels on YouTube, TikTok, Instagram, Twitter, etc.

They've also [said](#):

We're particularly interested in funding massively scalable projects: projects that could scale up to productively spend tens or hundreds of millions of dollars per year.

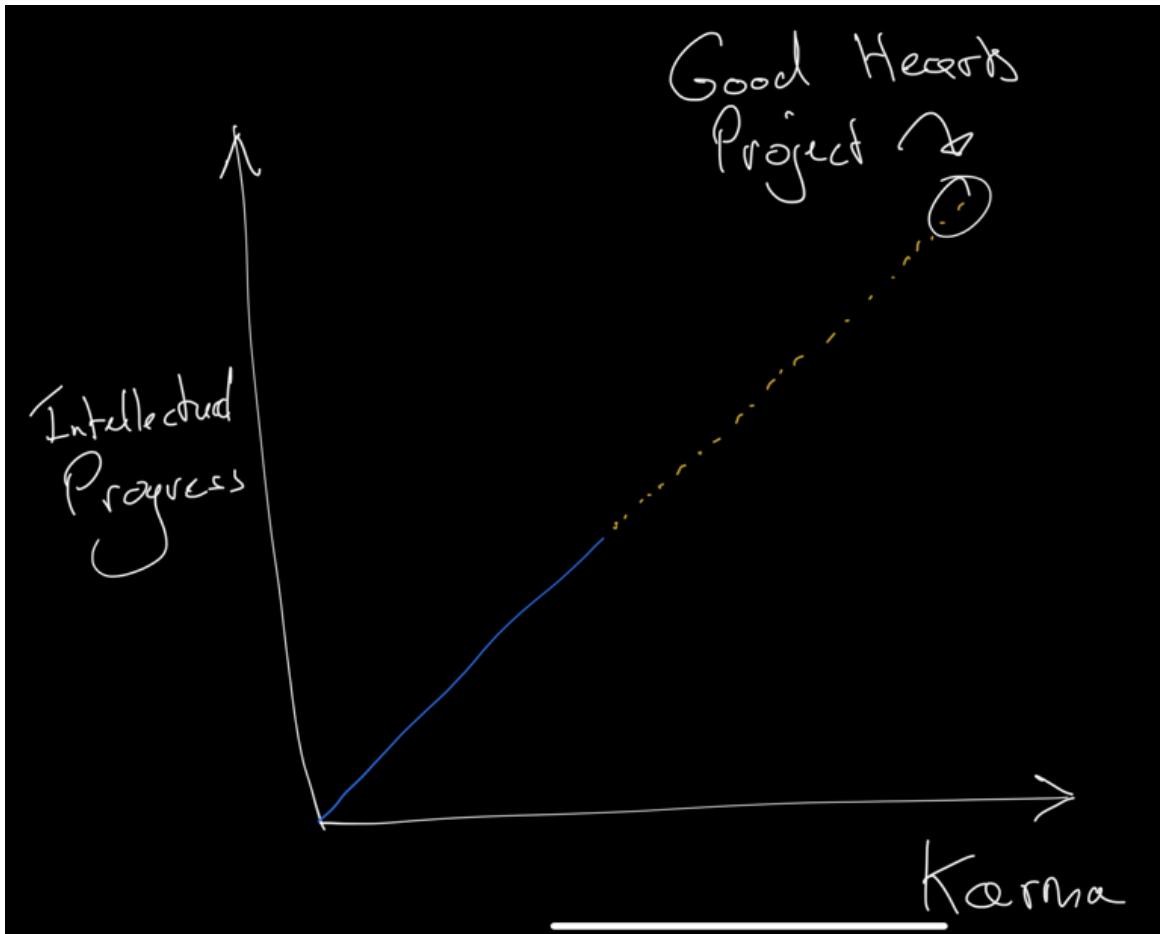
We are the best of both worlds: A blog that FTX and other funders can just *pour* money into. Right now we're trading \$1 per Good Heart Token, but in the future we could 10x or 100x this number and possibly see linear returns in quality content!

Trends Generally Continue Trending

Paul Christiano has [said](#):

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 be accurate is whether it's a trend extrapolation and how far in the future it is.

We agree with this position. So here is our trend-extrapolation argument, which we think has been true for many years and so will continue to be true for at least a few years.



The blue line is the observed trend, the yellow dotted line is the extrapolation.

So far it seems like higher-karma posts have been responsible for better insights about rationality and existential risk. The natural extrapolation suggests it will increase if people produce more content that gets high karma scores. Other trends are *possible*, but they're not *probable*. The prior should be that the trend continues!

However, epistemic modesty does compel me to take into account the possibility that we are wrong on this simple trend extrapolation. To remove any remaining likelihood of karma and intellectual progress becoming decoupled, **we would like to ask all users participating in the Good Hearts Project to really try hard to not be swayed by any unaligned incentives** (e.g. the desire to give your friends money).

Yes, We Have Taken User Feedback (GPT-10 Simulated)

We care about hearing arguments for and against our decision, so we consulted with a proprietary beta version of GPT-10 to infer what several LessWrong users would say about the potential downsides of this project. Here are some of the responses.

[Eliezer Yudkowsky](#): "To see your own creation have its soul turned into a monster before your eyes is a curious experience."

[Anna Salamon](#): "I can imagine a world where earnest and honest young people learn what's rewarded in this community is the most pointed nitpick possible under a post and that this might be a key factor in our inability to coordinate on preventing existential risk".

[Kaj Sotala](#): "I am worried it would lead to the total collapse of an ecosystem I've poured my heart and soul into in the past decade".

[Scott Garrabrant](#): "Now I can get more Good Heart tokens than Abram Demski! This is exactly what I come to LessWrong for."

[Clippy](#): "What's the exchange rate between these tokens and paperclips?"

[Quirinus Quirrell](#): "A thousand of your measly tokens pale in comparison to a single Quirrell point."

[Zvi Mowshowitz](#): "I would never sell my soul for such a low price. LessWrong *delenda est.*" (Then my computer burst into flame, which is why this post was two hours late.)

(Of course, this is only what our beta-GPT-10 simulations of these users said. The users are welcome to give their actual replies in the comment section below.)

Our EA funders have reviewed these concerns, and agree that there are risks, but think that, in the absence of anything better to do with money, it's worth trying to scale this system.

A Final Note on Why

I think the work on LessWrong matters a lot, and I'd like to see a world where these people and other people like them can devote themselves full-time to producing such work, and be financially supported when doing so.

Future Enhancement with Machine Learning

We're hoping to enhance this in the future by using machine learning on users' content to predict the karma you will get. Right now you only get Good Heart tokens *after* your content has been voted on, but with the right training, we expect our systems will be able to predict how many tokens you'll receive *in advance*.

This will initially look like people getting the money for their post at the moment of publishing. Then it will look like people getting the money when they've opened their draft and entered the title. Eventually, we hope to start paying people the moment they create their accounts.

For example, Terence Tao will create a LessWrong account, receive \$10 million within seconds, and immediately retire from academia.

Good Hearts Laws

- While we're rolling it out, only votes from existing accounts will count, and only votes on new content will count. No, self-votes will not count.
- There is currently a cap of 600 Good Heart Tokens per user while we are rolling out the system.
- The minimum number of tokens to be exchanged is 25. If you don't make that many tokens, we will not take the time to process your request. (Go big or go home.)

"Together, people with good hearts [tokens] and fine minds can move the world."

— **Bill Clinton** [3]

1. ^

<https://libquotes.com/dalai-lama/quote/lbs9z8y>

2. ^

<http://www.literaturepage.com/read/shakespeare-henry-iv-part2-33.html>

3. ^

<https://www.azquotes.com/quote/935663>

Call For Distillers

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

Many technical alignment researchers are bad-to-mediocre at writing up their ideas and results in a form intelligible to other people. And even for those who are reasonably good at it, writing up a good intuitive explanation still takes a lot of work, and that work lengthens the turn-time on publishing new results. For instance, a couple months ago I wrote [a post](#) which formalized the idea of abstractions as redundant information, and argued that it's equivalent to abstractions as information relevant at a distance. That post came out about two months after I had the rough math worked out, because it took a lot of work to explain it decently - and I don't even think the end result was all that good an explanation! And I *still* don't have a post which explains well why that result is interesting.

I think there's a lot of potential space in the field for people who are good at figuring out what other researchers' math is saying intuitively, and why it's interesting, and then communicating that clearly - i.e. the skill of *distillation*. This post will briefly sketch out what two kinds of distillation roles might look like, what skills are needed, and talk about how one might get started in such a role.

Two Distiller Roles

The two types of distiller role I'll sketch are:

- “Independent” distiller: someone who works independently, understanding work published by other researchers and producing distillations of that work.
- “Adjunct” distiller: someone who works directly with one researcher or a small team, producing regular write-ups of what the person/team is thinking about and why.

These two roles add value in slightly different ways.

An independent distiller's main value-adds are:

- Explaining the motivation and intended applications
- Coming up with new examples
- Boiling down the “key intuitive story” behind an argument
- Showing how the intuitive story fits into the context of the intended applications

I expect the ability to come up with novel examples and boil down the core intuitive story behind a bunch of math are the rate-limiting skills here.

[Rob Miles](#) is a good example of an existing independent distiller in the field. He makes YouTube videos intuitively explaining various technical results and arguments. Rob's work is aimed somewhat more at a popular audience than what I have in mind, but it's nonetheless been useful for people in the field.

I expect an adjunct distiller's main value-adds are:

- Writing up explanations, examples, and intuitions, similar to the independent distiller
- Saving time for the technical researcher/team; allow more specialization
- Providing more external visibility/legibility into the research process and motivation
- Accelerating the research process directly by coming up with good examples and intuitive explanations

I expect finding a researcher/team to work with is the rate-limiting step to this sort of work.

[Mark Xu](#) is a good example of an existing adjunct distiller. He's worked with both Evan Hubinger and Paul Christiano, and has written up decent distillations of some of their thoughts. I believe Mark did this with the aim of later doing technical research himself, rather than mostly being a distiller. That is a pretty good strategy and I expect it to be a common pathway, though naturally I expect people who aim to specialize in distillation long-term will end up better at distillation.

What Kind Of Skills Are Needed?

I expect the key rate-limiting skills are:

- Ability to independently generate intuitive examples when reading mathematical arguments, or having a mathematical discussion
- Ability to extract the core intuitive story from a mathematical argument
- Writing/drawing skills to clearly convey technical intuitions to a wider audience
- Ability to do most of the work of [crossing the communication gap](#) yourself - both so that researchers do not need to spend a lot of effort communicating to you, and so that readers do not need to spend a lot of effort understanding you
- For the adjunct role, [ability to write decent things quickly and frequently without too much perfectionism](#)
- For the non-adjunct role, ability to do all this relatively independently

How To Get Started

Getting started in an independent distiller role should be pretty straightforward: choose some research, and produce some distillations. It's inherently a very legible job, so you should pretty quickly have some good example pieces which you could showcase in a grant application (e.g. from the [Long Term Future Fund](#) or [FTX Future Fund](#)). That said, bear in mind that you may need some practice before you actually start to produce very good distillations.

An adjunct role is more difficult, because you need someone to work with. Obvious advice: **just asking people is an underutilized strategy, and works surprisingly well**. Be sure to emphasize your intended value-add to the researcher(s). If you want to prove yourself a bit before reaching out, independently distilling some of a researcher's existing public work is another obvious step. You might also try interviewing a researcher on some part of their work, and then distilling that, in order to get a better feel for what it would be like to work together before actually committing.

A concrete bet offer to those with short AI timelines

[Update 2 (11/4/2022): Matthew Barnett now thinks he will probably lose the 2026 bet (>50% credence).]

[Update 1: we have taken this bet with two people, as detailed in [a comment below](#).]

Recently, [a post claimed](#),

it seems very possible (>30%) that we are now in the crunch-time section of a short-timelines world, and that we have 3-7 years until Moore's law and organizational prioritization put these systems at extremely dangerous levels of capability.

We (Tamay Besiroglu and I) think this claim is strongly overstated, and disagree with the suggestion that "It's time for EA leadership to pull the short-timelines fire alarm." This post received a fair amount of attention, and we are concerned about a view of the type expounded in the post causing EA leadership to try something hasty and ill-considered.

To counterbalance this view, we express our disagreement with the post. To substantiate and make concrete our disagreement, **we are offering to bet up to \$1000 against the idea that we are in the “crunch-time section of a short-timelines”.**

In particular, we are willing to bet at 1:1 odds that no more than one of the following events will occur by 2026-01-01, or alternatively, 3:1 odds (in our favor) that no more than one of the following events will occur by 2030-01-01.

- A model/ensemble of models achieves >80% on all tasks in the [MMLU](#) benchmark
- A credible estimate reveals that an AI lab deployed EITHER $>10^{30}$ FLOPs OR hardware that would cost \$1bn if purchased through competitive cloud computing vendors at the time on a training run to develop a single ML model (excluding autonomous driving efforts)
- A model/ensemble of models will achieve >90% on the [MATH dataset](#) using a no-calculator rule
- A model/ensemble of models achieves >80% top-1 strict accuracy on competition-level problems on the [APPS benchmark](#)
- A gold medal for the IMO Grand Challenge (conditional on it being clear that the questions were not in the training set)
- A robot that can, from beginning to end, reliably wash dishes, take them out of an ordinary dishwasher and stack them into a cabinet, without breaking any dishes, and at a comparable speed to humans (<120% the average time)
- Tesla's full-self-driving capability makes fewer than one major mistake per 100,000 miles

Since we recognize that betting incentives can be weak over long time-horizons, we are also offering the option of employing Tamay's [recently described betting procedure](#) in which we would enter a series of repeated 2-year contracts until the resolution date.

Specific criteria for bet resolution

For each task listed above, we offer the following concrete resolution criteria.

Task	Full details
A model/ensemble of models achieves >80% on all tasks in the MMLU benchmark	A public document reveals that a model or ensemble of models had an average accuracy score of more than 80.0% on every task listed in the Measuring Massive Multitask Language Understanding benchmark by Dan Hendrycks et al.
A credible estimate reveals that an AI lab deployed EITHER $>10^{30}$ FLOPs OR hardware that would cost \$1bn if purchased through competitive cloud computing vendors at the time on a training run to develop a single ML model (excluding autonomous driving efforts)	If there is a dispute, our counterparty will offer an example that they believe may trigger this condition to resolve positively. Then, we will use the method described in this post to estimate the number of FLOPs used for a given training run. We will then find the competitive rates by taking the <i>minimum</i> of the most recent prices found on Google Cloud, Microsoft Azure, and Amazon AWS, or the equivalent services corresponding to each of those companies if their names change in the future. We will determine the rate corresponding to the processing unit that would be capable of training the model in the paper. If the training run was completed than 2 years before resolution, an adjustment is applied to the cost of the training run to account for price-declines in computing. This adjustment will be based on a credible estimate of price-performance declines in the last few years, OR, a constant decline of 30% per year, whichever is more favorable to the counterparty. We will use inflation-adjusted 2022 US dollars, according to the CPI according to the BLS .
A model/ensemble of models will achieve >90% on the MATH dataset using a no-calculator rule.	A public document reveals that a model or ensemble of models had an average accuracy score of more than 90% when evaluated on the MATH benchmark by Dan Hendrycks et al. The no-calculator rule ensures that competitors don't use calculators or script executors. As Hendrycks explains , "With a script executor, competitors do not need to figure out how to succinctly reason to the conclusion and cleverness is rarely needed."
A model/ensemble of models achieves >80% top-1 strict accuracy on competition-level problems on the APPS benchmark	A public document reveals that a model or ensemble of models had an average top-1 strict accuracy score of more than 80% on the competition-level coding problems listed in the APPS benchmark by Dan Hendrycks et al. Top-1 accuracy is distinguished, as in the paper, from top-k accuracy in which k outputs from the model are generated, and the best output is selected. Strict accuracy "requires programs pass every test case" (as defined in section 4.2 of the paper).

A gold medal for the IMO Grand Challenge	We will use the same criteria as revealed by Paul Christiano here . To clarify, we will not require that the model be open-source, only that it clearly demonstrates the ability to obtain a gold medal for the IMO.
A robot that can, from beginning to end, reliably wash dishes, take them out of an ordinary dishwasher and stack them into a cabinet, without breaking any dishes, and at a comparable speed to humans (<120% the average time)	For this task, a credible video or report must be public that clearly demonstrates that this ability is feasible to a very high level of confidence. Videos or reports that do not report rates of system malfunctions, or do not permit the calculation of how much time it would take to take out and stack at least 10 dishes, are not admissible. An error is something that would have a reasonable chance of a dish being broken if performed in an ordinary setting. Reliability is defined as an error rate below 1 percent across all full loading-and-unloading tests, when given natural starting setups ie. pre-programmed precise robotic planning is not admissible. This blog post will be consulted to determine the average time it takes to load and unload dishes, under the section "Time Study Of Dishes". In case there is a dispute over this question, our counterparty will offer an example that they believe may trigger this condition to resolve positively. We will then email the first author of whatever paper, blog post, or video in question asking whether they believe their system satisfies the conditions specified here, using their best judgment.
Tesla's full-self-driving capability makes fewer than one major mistake per 100,000 miles	We will require that a credible document from Tesla report crash or error-rate statistics for cars using their full self-driving mode . The error rate must include errors that would have happened if it were not for the presence of human intervention and it must be calculated over a reasonable distribution over all weather and road conditions that humans normally drive in. The cars must be able to go from arbitrary legal locations in the US to other arbitrary locations without errors or interventions, at least in theory. A "major mistake" is defined as a mistake that requires human intervention to correct, as otherwise the car would grossly violate a traffic law, crash into an object, prevent the car from getting to its intended destination, or have the potential to put human life into serious jeopardy.

Some clarifications

For each benchmark, we will exclude results that employed some degree of cheating. Cheating includes cases in which the rules specified in the original benchmark paper are not followed, or cases where some of the test examples were included in the training set.

Moses and the Class Struggle

"Take off your sandals. For you stand on holy ground," said the bush.

"No," said Moses.

"Why not?" said the bush.

"I am a Jew. If there's one thing I know about this universe it's that there's no such thing as God," said Moses.

"You don't need to be certain I exist. It's a trivial case of Pascal's Wager," said the bush.

"Who is Pascal?" said Moses.

"It makes sense if you are beyond time, as I am," said the bush.

"Mysterious answers are not answers," said Moses.

"Take off your shoes and I will give you the power of God. Surely that is a profitable bet even if there is a mere 1% chance I exist," said the bush.

"It's a profitable bet if there is a mere 0.001% chance you exist," said Moses.

"Are you 99.999% sure I don't exist," said the bush.

"No," said Moses.

"Then take off your sandals," said the bush.

"No," said Moses.

"Why not?" said the bush.

"Categorical imperative. If I accepted bets with large risk in exchange for large payoff then anyone could manipulate me just by promising a large payoff. I need at least some proof you're real," said Moses.

The bush burst into flames.

"That's supposed to convince me of divine power? I've seen fires before," said Moses.

"Reach your hand into the flames," said the bush.

Moses carefully examined the flames. The bush burned but did not seem to be harmed by the fire. Moses waved his staff through the fire. It burned too but emerged unharmed. He felt his staff. It remained cool to the touch. Moses placed his staff into the flames again, this time for longer. His staff caught fire like the bush but once again was unharmed. Moses quickly flicked his hand through the flames. Nothing happened. Moses rested his hand inside the flames. He felt the heat but it didn't harm his hand nor did it cause him pain. Moses retrieved his hand.

"Does that convince you I am real?" said the bush.

"Nothing you can say or do will convince me God is speaking to me because the odds of God being real are lower than the odds I have become schizophrenic," said Moses.

"That's the least rational thing I've ever heard. Rationality is about updating your beliefs in the face of evidence. You have just declared that no quantity of evidence will change your mind," said the bush.

"That's a Bayesian argument. I'm a Frequentist," said Moses. He waved his staff through the flames until $p < 0.05$ and then he took off his sandals.

Later, in Cairo.

"Brother! It is so good to see you again," said Ramesses.

"I wish I could say the same," said Moses.

"What's wrong?" said Ramesses.

"I don't really know how to explain this," said Moses.

"Just tell it to me straight. You know you can talk about anything with me. I've never judged you. I've never even gotten mad. I've always been on your side and I always will be," said Ramesses.

Moses took a deep breath. He let it out slowly. "I talked to God. He says to let His people go."

Ramesses stared blank for a moment. Then he laughed so hard he fell over into the cushions. "You had me going there for a minute. You always were the jokester. I missed you so much. Nobody makes fun of me anymore since you left. I don't blame them. It's dangerous to tease someone who can execute you for treason, insubordination, heresy or just pure whim. I know I have nothing to complain about. I live an incredibly privileged life. But privilege comes with its price. You're a breath of fresh air in an ocean of sycophants."

"This isn't a joke. I'm serious," said Moses.

"And my favorite thing about you is how you commit to the part. Like that joke where you pretended to be the son of an Israelite. I still can't believe you actually got yourself circumcised. Mother was so furious," said Ramesses.

"Goddammit," said Moses, "Ow!"

"What?" said Ramesses.

"My staff. I got a splinter," said Moses. He tried to pull it out with his fingernails but it didn't work.

"Be delicate with your words while in the presence of the avatar of Ra," said Ramesses melodramatically, "You might offend God."

"Now you're messing with me," said Moses. He put his hand in his mouth and tried to remove the splinter with his teeth.

"Maybe you can fool the plebs—or even the high priests. But pretending to be God is literally my full-time job. I know exactly what goes on behind the curtain," Ramesses gestured at workers who were covering a giant pyramid with polished white limestone, "Look there. What do you see?"

"I see an enslaved nation toiling away for the vanity of a false god," said Moses. His teeth caught on the splinter and he finally yanked it out.

"Then you are factually wrong. Slaves work in the fields. Those masons are contractors. Skilled craftsmen. Centuries from now, future civilizations will dig up these construction sites and find animal bones proving these employees and small business owners ate meat. This will be our legacy. The birth of a middle class," said Ramesses.

"You're paying people to build a giant stone triangle?" said Moses, "Why?"

"To bring about a workers' paradise," said Ramesses.

"Back up," said Moses.

Ramses removed the bronze ankh from his neck and placed it in Moses' hands. "What do you see?"

"A competitive status good with no intrinsic value," said Moses.

"This item is pretty and therefore does possess intrinsic value. But that's beside the point. What do you think goes into making one of these?" said Ramesses.

"Copper and tin," said Moses.

"That copper comes from Sardinia, Cyprus and Tyrol. They are all located on the far side of the Mediterranean Sea. Rumor says the tin comes from Northwest Europe, but every good scholar knows England is a myth. The tin is actually from Iberia," said Ramesses.

"Hug the query. What does globalization have to do with giant stone triangles?" said Moses.

"I am getting to that. We do not just import copper and tin. Egypt has no natural deposits of lapis lazuli, silver or obsidian. We depend on the Hellenistic city-states for mineral resources. They depend on our grain exports. Civilization is fragile. One major disaster and the Bronze Age is over. Two million people starve to death in Egypt alone," said Ramesses.

"Better two million Egyptians starve today than ten million Egyptians starve when your dynasty falls," said Moses.

"We don't have to make that choice if we can increase production sufficiently," said Ramesses.

"A post-scarcity society built on the backs of slaves is a dystopia," said Moses.

"Slavery is just a transitory phase. I am building an industrial consumer economy. Capital has the potential to compound faster than humans breed," said Ramesses.

"That sounds unsustainable," said Moses.

"It is unsustainable. Once enough wealth is accumulated, the proletariat will overthrow the bourgeoisie and establish a workers' paradise," said Ramesses.

"Justice delayed is justice denied," said Moses.

"Shut up and multiply," said Ramesses.

"You still haven't answered the question about why you're building a giant stone triangle," said Moses.

"Initial industrial capacity is a product of government-stimulated demand," said Ramesses.

"You could use the same argument to build walls or canals. What possible use could a giant stone triangle have?" said Moses.

"To live forever," said Ramesses.

"You never seemed interested in monuments when we were growing up. Perhaps I missed something," said Moses.

"It is not about establishing a legacy. I am literally going to live forever. The high priests will preserve my kidneys, heart and lungs. In a hundred years or so, the post-scarcity communist utopia will import water from the Fountain of Youth and they will bring me back to life," said Ramesses.

"What about your head?" said Moses.

"My skull is preserved too. After cleaning it, of course. It's incredible how much of a dead brain can be removed via the nostrils without damaging the skull," said Ramesses.

"Your plan is never going to work. There are too many places for it to go wrong. What if someone breaks into your tomb and steals your skull? They might use it for evil," said Moses.

"Nobody could establish a communist state without having solved the human coordination problem. Thus, any future civilization with the magic to evade my traps and break into my tomb would surely be benevolent," said Ramesses.

"Someone might wield tremendous magical power while disagreeing with you about an important issue," said Moses.

"Nonsense. Rational agents with common knowledge of each other's beliefs cannot agree to disagree," said Ramesses.

Moses' staff twitched.

dalle2 comments

(i.)

On April 6, OpenAI announced “DALLE-2” to great fanfare.

There are two different ways that OpenAI talks to the public about its models: as research, and as marketable products.

These two ways of speaking are utterly distinct, and seemingly immiscible. The OpenAI blog contains various posts written in each of the two voices, the “research” voice and the “product” voice, but each post is wholly one or the other.

Here’s OpenAI introducing a model ([the original DALLE](#)) in the research voice:

DALL·E is a 12-billion parameter version of GPT-3 trained to generate images from text descriptions, using a dataset of text-image pairs. [...]

Like GPT-3, DALL·E is a transformer language model. It receives both the text and the image as a single stream of data containing up to 1280 tokens, and is trained using maximum likelihood to generate all of the tokens, one after another. This training procedure allows DALL·E to not only generate an image from scratch, but also to regenerate any rectangular region of an existing image that extends to the bottom-right corner, in a way that is consistent with the text prompt.

And here they are, introducing a model ([Codex](#)) in the product voice:

OpenAI Codex is a descendant of GPT-3; its training data contains both natural language and billions of lines of source code from publicly available sources, including code in public GitHub repositories. [...]

GPT-3’s main skill is generating natural language in response to a natural language prompt, meaning the only way it affects the world is through the mind of the reader. OpenAI Codex has much of the natural language understanding of GPT-3, but it produces working code—meaning you can issue commands in English to any piece of software with an API. OpenAI Codex empowers computers to better understand people’s intent, which can empower everyone to do more with computers.

Interestingly, when OpenAI is planning to announce a model as a product, they tend not to publicize the research leading up to it.

They don’t write research posts as they go along, and then write a product post once they’re done. They *do* publish the research, in PDFs on the Arxiv. The researchers who were directly involved might tweet a link to it. And of course these papers are immediately noticed by ML geeks with RSS feeds, and then by their entire social networks. It’s not like OpenAI is trying to *hide* these publications. It’s just not making a big deal out of them.

Remember when the GPT-3 paper came out? It didn’t get a splashy announcement. It didn’t get noted in the blog at all. It was just dumped unceremoniously onto the Arxiv.

And then, a few weeks later, they donned the product voice and announced the “[OpenAI API](#),” GPT-3 as a service. Their post on the API was full of enthusiasm, but contained almost no technical details. It mentioned the term “GPT-3” only in passing:

Today the API runs models with weights from the GPT-3 family with many speed and throughput improvements. Machine learning is moving very fast, and we’re constantly upgrading our technology so that our users stay up to date.

It didn’t even mention how many parameters the model had!

(My tone might sound negative. Just to be clear, I’m not trying to criticize OpenAI for the above. I’m just pointing out recurring patterns in their PR.)

(ii.)

DALLE-2 was announced in the product voice.

In fact, their [blog post](#) on it is the most over-the-top product-voice-y thing imaginable. It goes so far in the direction of “aiming for a non-technical audience” that it ends up seemingly addressed to *small children*. As in a picture book, it offers simple sentences in gigantic letters, one or two per page, each one nestled between huge comforting expanses of blank space. And sentences themselves are ... well, stuff like this:

DALL·E 2 can create original, realistic images and art from a text description. It can combine concepts, attributes, and styles. [...]

DALL·E 2 has learned the relationship between images and the text used to describe them. [...]

Our hope is that DALL·E 2 will empower people to express themselves creatively.

It’s honestly one of the weirdest, most offputting, least appealing web pages I’ve seen in a long time. But then, I’m clearly not in its target demographic. I have no idea who the post *is* intended for, but whoever they are, I guess maybe this is the kind of thing they like? IDK.

Anyway, weird blog post aside, what *is* this thing?

The examples we’ve seen from it are pretty impressive. OpenAI clearly believes it’s good enough to be a viable product, the pictorial equivalent of GPT-3. So it must be doing something new and different, right – behind the scenes, in the research? Otherwise, how could it be better than anything else out there?

There’s also an associated [research paper](#) on the Arxiv.

I asked myself, “how is DALLE-2 so good?”, and then read the research paper, naively imagining it would contain a clear answer. But the paper left me confused. It *did* introduce a new idea, but it was kind of a weird idea, and I didn’t quite believe that *it*, of all things, could be the answer.

It turns out that DALLE-2 is a combination of 3 things:

1. **GLIDE**, a model from an earlier (Jan 2022) [OpenAI paper](#).

The GLIDE paper was released around the same time as the (GLIDE-like) image models I made for [@nostalgebraist-autoresponder](#), and I talked about it in [my post on those](#).

2. **An additional upsampling model** that converts GLIDE's 256x256 images into larger 1024x1024 images.

This one is pretty trivial to do (if compute-intensive), but the bigger images are probably responsible for some of DALLE-2's "wow effect."

3. **"unCLIP"**: a novel, outside-the-box, and (to me) counterintuitive architecture for conditioning an image model on text.

The only part that's really new is unCLIP. Otherwise, it's basically just GLIDE, which they published a few months ago. So, is unCLIP the magical part? Should I interpret DALLE-2 as tour de force demo of unCLIP's power?

Well, maybe not.

Remember, when OpenAI is building up to a "product voice" release, they don't make much noise about the research results leading up to it. Things started to make a lot more sense when I realized that "GLIDE" and "DALLE-2" are not two different things. They're clearly products of the same research project. Inside OpenAI, there'd be no reason to draw a distinction between them.

So maybe the magic was already there in GLIDE, the whole time.

Indeed, this seems very plausible. GLIDE did two new things: it added text conditioning, *and it scaled up the model*.

The text conditioning was presented as the main contribution of the paper, but the scaling part was not a small change: GLIDE's base image model is 2.3B parameters, around 8 times bigger than its analogue in their [mid-2021 diffusion paper](#), which was state-of-the-art at the time.

(*Sidenote: the base model in the [cascaded diffusion paper](#) was pretty big too, though - around half the size of GLIDE, if I'm doing the math right.*)

I'd guess that much of the reason DALLE-2 is so visually impressive is just that it's bigger. This wasn't initially obvious, because it's "only" as big as GLIDE, and size wasn't "supposed to be the point" of GLIDE. But the scale-up *did* happen, in any case.

(*Sidenote: OA made the unusual choice to release a dramatically weakened version of GLIDE, rather than releasing the full model, or releasing nothing. After training a big model, they released a structurally similar but tiny one, smaller than my own bot's base model by half. As if that wasn't enough, they also applied a broad-brush policy of content moderation to its training data, e.g. ensuring it can't generate porn by preventing it from ever seeing any picture of any human being.*

So you can "download GLIDE," sort of, and generate your own pictures with it ... but they kind of suck. It's almost like they went out of their way to make GLIDE feel boring, since we weren't supposed to be excited until the big planned release day.)

But what fraction of the magic is due to scale, then, and what fraction is due to unCLIP? Well, what exactly *is* unCLIP?

(iii.)

unCLIP is a new way of communicating a textual description to an image-generating model.

I went into more detail on this in that [post about the bot](#), but in short, there are multiple ways to do this. In some of them (like my bot's), the model can actually “see the text” as a sequence of individual words/tokens. In other ones, you try to compress all the relevant meaning in the text down into a single fixed-length vector, and that's what the model sees.

GLIDE actually did both of these at once. It had a part that processed the text, and it produced two representations of it: a single fixed-length vector summary, and another token-by-token sequence of vectors. These two things were fed into different parts of the image model.

All of that was learned from scratch during the process of training GLIDE. So, at the end of its training, GLIDE contained a model that was pretty good at reading text and distilling the relevant meaning into a single fixed-length vector (among other things).

But as it happens, there was already an existing OpenAI model that was *extremely* good at reading text, and distilling the relevant meaning into a single fixed-length vector. Namely, CLIP.

CLIP doesn't generate any pictures or text on its own. Only vectors. It has two parts: one of them reads text and spits out a vector, and the other one looks at an image and spits out another vector.

During training, CLIP was encouraged to make these vectors close to one another (in cosine similarity) if the text “matched” the image, and far from one another if they didn't “match.” CLIP is very good at this task.

And to be good at it, it has to be able to pack a lot of detailed meaning into its vectors. This isn't easy! CLIP's task is to invent a notation system that can express the essence of (1) *any possible picture*, and (2) *any possible description of a picture*, in only a brief list of maybe 512 or 1024 floating-point numbers. Somehow it does pretty well. A thousand numbers are enough to say a whole lot about a picture, apparently.

If CLIP's vectors are so great, can't we just use *them* to let our image models understand text? Yes, of course we can.

The most famous way is “*CLIP guidance*,” which actually still doesn't let the image model *itself* glimpse anything about the text. Instead, it tells the text-vector side of CLIP to distill the text into a vector, and then repeatedly shows the gestating image to the image-vector side of CLIP, while it's still being generated.

The image-vector side of CLIP makes a vector from what it sees, checks it against the text vector, and says “hmm, these would match more if the picture were a little more like *this*.” And you let these signals steer the generation process around.

If you want to let the image model actually *know* about the text, and learn to use it directly, rather than relying on CLIP's backseat driving ... well, CLIP gives you an easy way to do that, too.

You just take the text-vector side of CLIP (ignoring the other half), use it to turn text into vectors, and feed the vectors into your image generator somewhere. Then train the model with this setup in place.

This is "[CLIP conditioning](#)." It's the same thing as the fixed-vector part of GLIDE's conditioning mechanism, except instead of training a model from scratch to make the vectors, we just use the already trained CLIP.

And like, that's the natural way to use CLIP to do this, right? CLIP's text-vector part can distill everything relevant about a caption into a vector, or it sure feels that way, anyway. If your image model can see these vectors, what more could it need?

But if you're OpenAI, at this point, you don't stop there. You ask a weird question.

What if we trained a model to turn CLIP's "text vectors" *into* CLIP's "image vectors"?

And then, instead of feeding in the text vectors directly, we added an intermediate step where we "turn them into" image vectors, and feed *those* in, instead?

That is, suppose you have some caption like "vibrant portrait painting of Salvador Dalí with a robotic half face." OK, feed that caption through CLIP's text-vector part. Now you have a "CLIP text vector."

Your *new* model reads this vector, and tries to generate a "CLIP image vector": the sort of vector that CLIP's image-vector would create, if you showed it an *actual* vibrant portrait painting of Salvador Dalí with a robotic half face.

And then *this* vector is what your image generator gets to see.

At first, this struck me as almost nonsensical. "CLIP text vectors" and "CLIP image vectors" aren't two different species. They're vectors in the same space - you can take cosine similarities between them!

And anyway, the entire thing CLIP is trained to do is make its "text vectors" as close as possible to its "image vectors." So, given a text vector, surely the "corresponding image vector" is just ... the same vector, right? That's the best possible match! But if a perfect score is that easy to get, then exactly what task is this new neural net supposed to be *doing*?

It made more sense after I thought about it some more, though I still don't really get why it works.

CLIP's task isn't *just* to make the text and image vectors as close as possible. Getting a perfect score on that part is easy, too: just spit out the *same* vector for all texts and all images.

However, CLIP also has to make the vectors *dissimilar* when there *isn't* a match. This forces it to actually use the vectors as a medium of communication, expressing features that can distinct images and texts from one another. Also, the *amount* of vector similarity has to capture its probabilistic credences about different conceivable matches.

For these reasons, CLIP generally won't try for a perfect match between the two vectors. Among other things, that behavior would make it unable to express different levels of confidence.

The image vectors aren't supposed to encode exactly one "ideal" caption, but a bunch of information that could affect the relevance of many different captions, to different extents. Likewise, the text vectors aren't encoding a single "ideal" image, they're encoding a bunch of information (...etc). Merely-partial matches are a feature, not a bug; they encode rational uncertainty about alternatives.

So we shouldn't be thinking about identical vectors. We should be thinking, roughly, about sets of vectors that are roughly some specific *level* of closeness to another vector (whatever suits the model's purposes). These look like spherical caps or segments.

These are the regions in which CLIP vectors can vary, without impacting the model's decisions at all. Its indifference curves, if you will. To CLIP, all the vectors in such a region "look" identical. They all match their counterparts to an equal extent, and that's all CLIP cares about.

In high dimensions, there's a lot of room everywhere, so these regions are big. Given a caption, there are many different image vectors CLIP would consider equally consistent with it.

This means that something is missing from CLIP! There's information it doesn't know. The information is implicit in CLIP's *training data*, but never made it from there into CLIP's weights.

Because not all images (or captions) are equally plausible, just considered on their own. CLIP knows nothing about plausibility. It doesn't know whether an image "looks real" or not, just on its own.

CLIP's regions of indifference contain many image vectors corresponding to real-looking images, but they could well contain many *other* image vectors whose images look like garbage, or like nothing. It would have no idea anything was wrong with these, and would happily go along matching them (nonsensically) with captions. Why should it care? In training, it only saw real images.

So the meaning of a CLIP text vector is not, "any image that matches me deserves my caption." The meaning is, "an image that matches me deserves my caption, *provided it is a real image you'd find somewhere online.*"

But when you're trying to generate images, you don't get to make that assumption.

I mean, you sort of *do*, though? That's what your image generator is being trained to do, apart from the conditioning aspect: it's learning to make plausible images.

So you'd hope that "CLIP text vector + general prior for plausible images" would be enough. The CLIP vector encodes a bunch of plausible matching images, and a bunch of other implausible ones, but your image model will naturally go for the plausible ones, right? I'm still confused about this, actually.

But anyway, OpenAI decided to deal with this by training a model to take in text vectors, and spit out image vectors. Obviously, there's more than one possible right answer, so your model will be able to generate many different possibilities pseudo-randomly, like GPT.

But it will specifically try to pick only *plausible* image vectors. Its outputs will not uniformly fill up those spherical caps and sections. They'll cluster up in patches of plausibility, like real images do.

OA calls this new model "unCLIP," because it sort of inverts CLIP, in a sense. (They also call it a "prior," which confused me until I realized it was expressing a *prior probability over images*, and that this was its value-add.)

And then your image generator sees these vectors, each of which refers to some *specific picture* (up to CLIP's own ability to tell two pictures apart, which has its limits). Whereas, the text vectors actually refer to entire *indifference sets* of images, some of which happen to be garbage.

(iv.)

What does this get us, practically speaking? Apparently - for some reason - it makes the diversity-fidelity tradeoff disappear!

What's the diversity-fidelity tradeoff? OK, a little more exposition.

These days, with this type of model, everyone uses a technique called "classifier-free guidance." Without going into detail, this basically takes a model that's already conditional, and makes it like ... *more* conditional.

Your model is trying do two things: generate some plausible picture, and generate a picture that specifically matches the text (or category, or whatever) that you fed in. Classifier-free guidance tells it, "you think that's a match? come on, make them match even more." And you can do more or less of it. It's a dial you can turn.

Usually, when you turn the dial way up, the model has a better chance of generating the thing you asked for ("fidelity") - but at the cost of diversity. That is, with guidance turned way up, the images start to look less and less different, eventually converging on slight variants of a single picture. Which probably *will* be a version of what you requested. But if you want another one instead, you're out of luck.

But when you have unCLIP, you can ask it to psuedo-randomly generate *many* different image vectors from the same text vector. These encode diverse images, and then you're free to turn the guidance up as much as you want. (It's just the guidance for that *one vector*, and the vectors are all different.) That solves the problem, doesn't it?

Weirdly (?), that is *not* what OpenAI does! No, they have unCLIP generate a single fixed vector, and then turn guidance up, and generate many images from this same vector. And somehow ... these images are still diverse??

Here's a picture from the paper showing this:



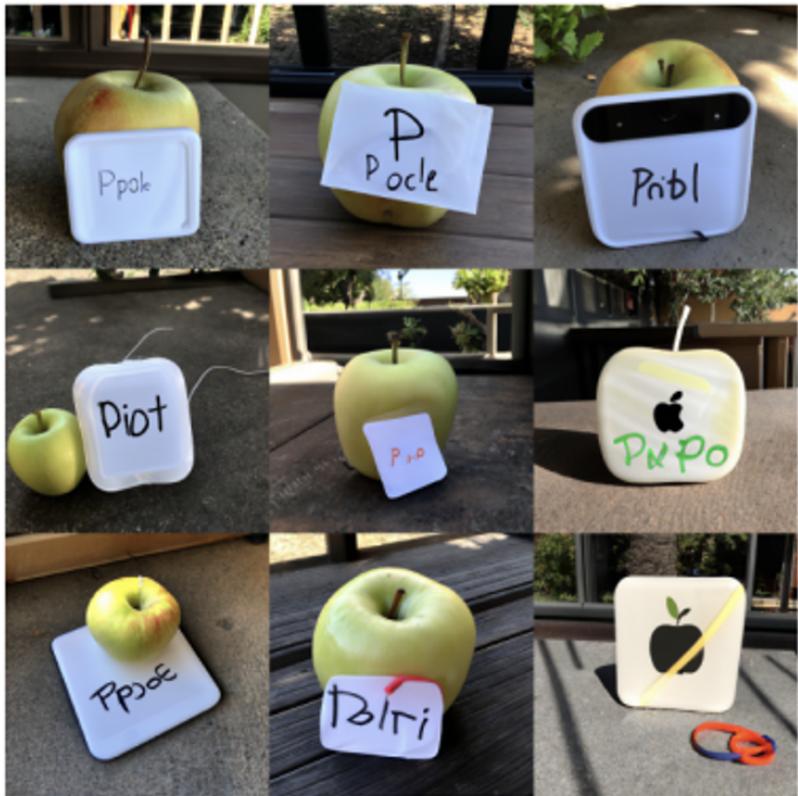
Figure 9: Samples when increasing guidance scale for both unCLIP and GLIDE, using the prompt, “A green vase filled with red roses sitting on top of table.” For unCLIP, we fix the latent vectors sampled from the prior,

The GLIDE panels display the classic diversity-fidelity tradeoff, with the POV and background becoming nearly identical at high guidance. But even if you guide hard on a single unCLIP vector, you still get vases in different-looking rooms and stuff. (*In fairness, it still doesn’t make them green, which GLIDE does.*)

This rides on the fact that CLIP cannot always fully tell two pictures apart, I guess. There are many different image that have the same CLIP image vector, so a single image vector can be “interpreted” in a variety of ways.

And indeed, the paper uses unCLIP to visualize this more directly, with fun galleries of images that all look the same to CLIP. (*Or close to the same; the generator might not do this perfectly.*)

Exactly what information *is* written in these CLIP image vectors? What do they say, if we could read them? Well, there’s one vector that expresses ... whatever all these pictures have in common:



If there's a concept there, it's an alien kind of concept. But you can see the family resemblance.

If I understand things correctly, the different apple/Apple-ish images above are precisely like the different vases from the earlier figure. They reflect the extent and nature of the diversity available given a single image vector.

But ... I still don't get why this is any better than conditioning on a single *text* vector. Like, if an image vector is ambiguous, surely a text vector is even more ambiguous? (It could map to many image vectors, each of which could map to many images.)

So why do models conditioned on a *text* vector (like GLIDE) collapse to a single image under conditioning, while a model conditioned on an *image* vector can still be diverse? Surely the text vectors aren't encoding the contents of one specific image *more precisely* than the image vectors do? The reverse should be true.

But the different is real, and they measure it quantitatively, too.

(v.)

Whence DALLE-2's magic?

TBH I think most of it is just in the scale. We're not used to seeing AI images this crisply detailed, this pretty, with this few glitches. I'd expect a bigger model to be superior in those ways.

The unCLIP thing ... well, I guess it lets them feel comfortable turning the guidance scale way up. And that raises the probability that you get what you asked for.

It's not that the model now has a *deeper understanding* of what you asked for, necessarily. (GLIDE could *do* the "green vase" prompt perfectly, it just couldn't do many variations of it.) What has changed is that you can now push the model really hard in the direction of *using* that understanding, while still getting varied results.

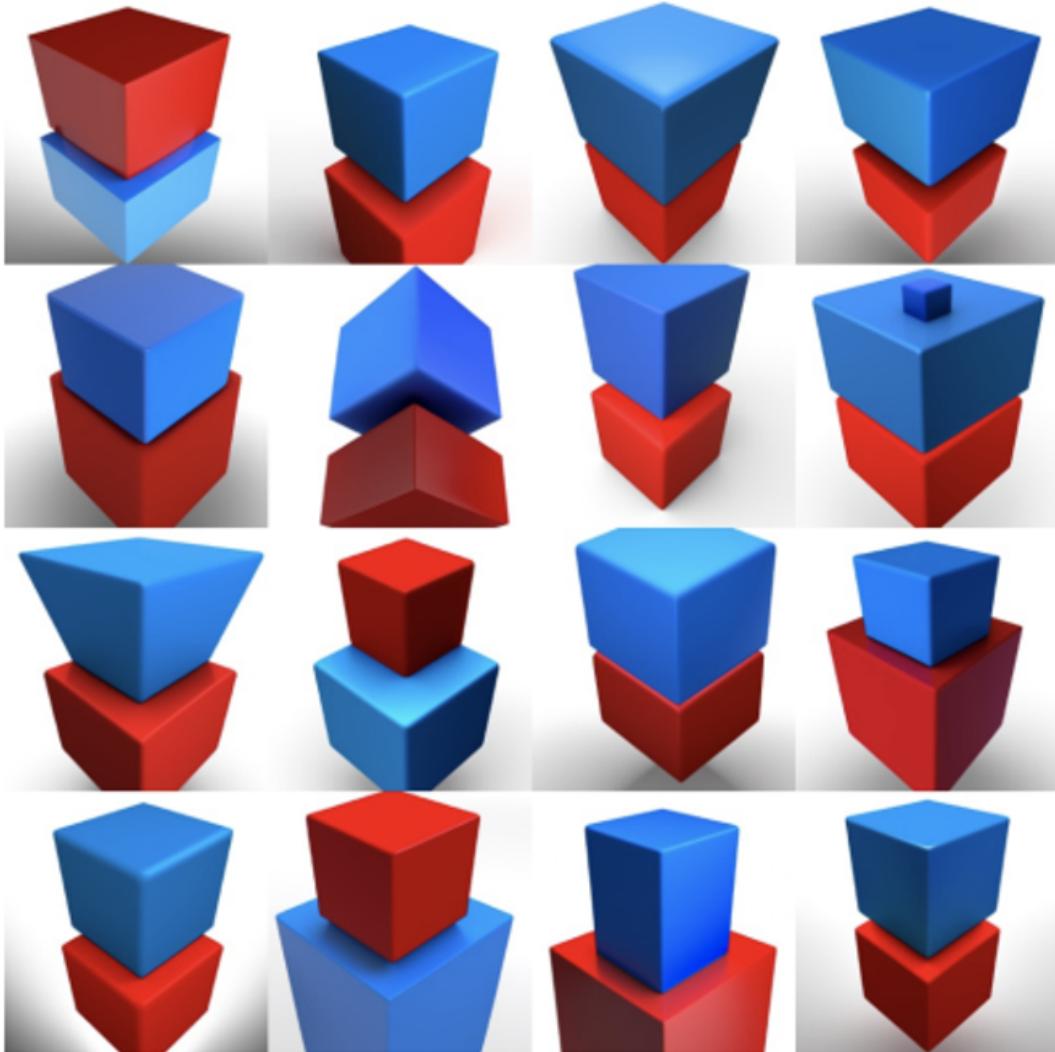
I think the single-vector style of conditioning is going to have some fundamental limits.

Along with the unCLIP vectors, OpenAI let their model see the other things GLIDE usually sees, from its own text-conditioning mechanism. But they talk about this as if it's sort of vestigial, and unCLIP is clearly sufficient/better.

But there are important distinctions that are not written into CLIP vectors. Is this a green apple, or an Apple device, or both? The CLIP vector can't tell you (see above).

In general, fixed-length vectors will struggle to express arbitrary layers of composition.

This is an issue for CLIP, which has trouble keeping track of which attributes go with which objects. Or how big anything is. Here's another batch of "identical" pictures:



In NLP, this kind of thing is what made attention + RNN models work so much better than pure RNNs, even before the transformer came along and dispensed with the RNN part.

If your model's inner representation of a text grows with the length of the text (GLIDE, etc), then when you feed it a longer string of composed clauses, it has correspondingly more "room" to record and process information about what's happening in them.

While a fixed-length vector (RNNs, CLIP) can only fit a fixed quantity of information, no matter how long the input.

But having GLIDE's sequential representation doesn't actually solve the problem. The model that produced these picture was looking at a prompt, with both the unCLIP and GLIDE lenses, and still it produced the pictures we see.

But I would be very surprised if it turned out that fixed-length vectors were *just as good*, in this context, as variable-length sequences. NLP is just not like that.

CLIP is very good, but we shouldn't resign ourselves to never besting it. There are things it struggles to represent which used to be hard for all NLP models, but no longer need to be hard.

It's odd that the GLIDE mechanism was so ineffective. But OpenAI (and everyone else) should keep trying things like that. CLIP is too limiting.

How good is DALL-E-2? As a reversed CLIP, it has all of CLIP's limitations.

Here is the paper's Figure 19, *Random samples from unCLIP for prompt "A close up of a handpalm with leaves growing from it."*

They are all very *pretty*, and hands and leaves are generally *involved*, but, well:



Slack gives you space to notice/reflect on subtle things

Followup to "[My slack budget: 3 surprise problems per week](#)"

Previously, I thought the reasons to preserve slack in your life (or in your organization) were to:

- Avoid using up all of your resources
- Avoid hitting a crisis where you suddenly have multiple surprise problems, and you have no choice but to either do a shitty job handling all of them, or deliberately not handle some of them at all (and deal with the consequences)
- Avoid feeling really unpleasant
- Allow you to live up to principles / be more pro-social.

This year in December/January, the Lightcone Infrastructure team (where I work) took on a large number of difficult projects at once. I was thinking about how wise/unwise this was, and chatting with John Wentworth about it. I listed the problems-with-lack-of-slack.

He said (something like): "Oh, that's not the point of slack. Or, not the part I'm most interested in. The point of slack is to give you the space to notice subtle things and think about them."

A rough model is something like, here are three types of things you can do:

1. You can take actions on whatever stuff it is you like to do on-purpose. (Typically your day job, or fun projects, or whatever)
2. You can rest/recover/do-random-fun-things.
3. You can... be cognizant of stuff going that isn't immediately relevant to the first two things, and mull it over, and notice new, potentially fruitful trains of thought about them.

When you're got too many things to do and stuff is constantly exploding and demanding your attention, #3 is the first thing to go. You often *need* to be putting out fires (#1), and if you do that too much and did into energy reserves your body will eventually be like "No, screw that, time to burn out for a bit and spend a week tired and recovering." (#2)

But, there won't be a moment where you experience a clear failure-and-control-mechanism that pushes you to spend time on #3. You just... won't notice a thing that you might have noticed.

And this is particularly important when you're working on problems you [don't understand how to solve](#) (such as AI alignment, or how to improve institutions, or learn/teach rationality, or, just, any ol' problem in your life you're currently confused about, or maybe haven't even yet realized that you have)

In the explore/exploit dichotomy, when solving a problem-you-don't-understand, having a train of thoughts in "explore" mode is pretty valuable. Slack gives you space for your shower thoughts to be in explore mode.

Everything I Need To Know About Takeoff Speeds I Learned From Air Conditioner Ratings On Amazon

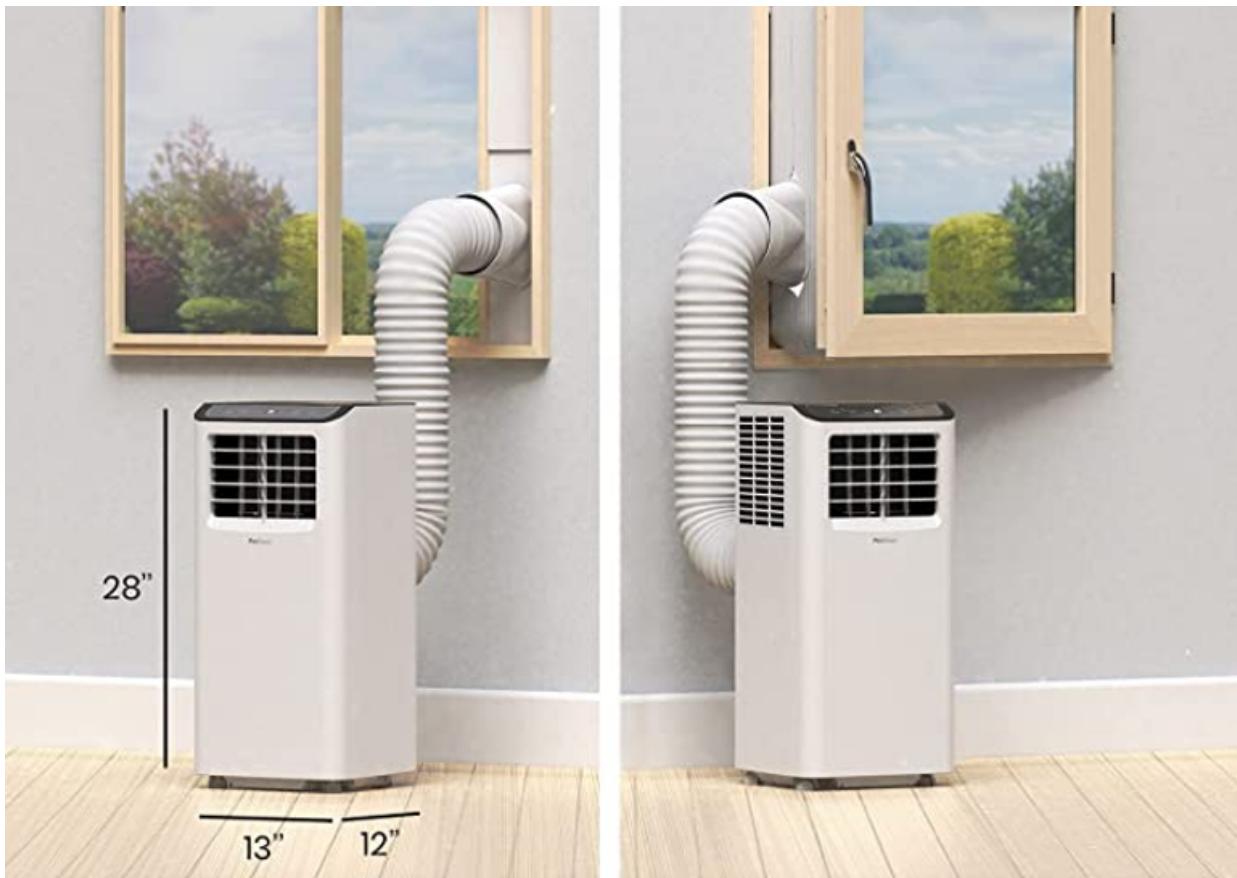
Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

I go to Amazon, search for “air conditioner”, and sort by average customer rating. There’s a couple pages of evaporative coolers (not what I’m looking for), one used window unit (?), and then [this](#):



Average rating: 4.7 out of 5 stars.

However, this air conditioner has a major problem. Take a look at this picture:



Window Kits Included.

Comes with a dual option window kit for slide and hinged windows.

Key thing to notice: there is one hose going to the window. Only one.

Why is that significant?

Here's how this air conditioner works. It sucks in some air from the room. It splits that air into two streams, and pumps heat from one stream to the other - making some air hotter, and some air cooler. The cool air, it blows back into the room. The hot air, it blows out the window.

See the problem yet?

Air is blowing out the window. In order for the room to not end up a vacuum, air has to come back *into* the room from outside. In practice, houses are very not airtight (we don't want to suffocate), so air from outside will be pulled in through lots of openings throughout the house. And presumably that air being pulled in from outside is *hot*; one typically does not use an air conditioner on cool days.

The actual effect of this air conditioner is to make the space right in front of the air conditioner nice and cool, but fill the rest of the house with hot outdoor air. Probably not what one wants from an air conditioner!

Ok, that's amusing, but the point of this post is not physics-101 level case studies in how not to build an air conditioner. The real fact of interest is that this is apparently the *top rated* new air conditioner on Amazon. How does such a bad design end up so popular?

One aspect of the story, presumably, is fake reviews. That phenomenon is itself a rich source of insight, but not the point of this post, and definitely not enough to account for the popularity of this air conditioner. The reviews shown on the product page are all "verified purchase", and mostly 5-stars. There are only 4 one-star reviews (out of 104). If most customers noticed how bad this air conditioner is, I do not think a 4.7 rating would be sustainable. Customers *actually do like* this air conditioner.

Customer reviews

★★★★★ 4.7 out of 5

104 global ratings

5 star 80%
4 star 15%
3 star 2%
2 star 0%
1 star 3%

▼ How are ratings calculated?

By feature

Feature	Rating
Remote Control	★★★★★ 4.7
Energy efficiency	★★★★★ 4.2
Noise level	★★★★★ 4.2

Review this product

Share your thoughts with other customers

Write a customer review

Top reviews from the United States

Jessica ★★★★★ R2D2
Reviewed in the United States on August 1, 2021
Verified Purchase
It's small and fits in all size places. Works great, doesn't push out air very strongly but the living room is no long humid
12 people found this helpful

Helpful | Report abuse

Julia ★★★★★ Surpassed expectations
Reviewed in the United States on November 14, 2021
Verified Purchase
Very powerful and will cool any bedroom in minutes. Also the size is compact compared to most other ac units. We have this ac unit in our computer room because our computers make the room temp very high and with this ac running, we feel very comfortable. We do connect this with a different outlet in another room because this unit does take up a lot of power and could trip a breaker but gosh does this make a room comfy in the summer time.
8 people found this helpful

Helpful | Report abuse

Maia Rodriguez ★★★★★ it works like it's intended to
Reviewed in the United States on July 30, 2021
Verified Purchase
I brought it for my family because my mom does not like the AC on. My family loves it
10 people found this helpful

Helpful | Report abuse

And hey, this air conditioner has a lot going for it! There's wheels on the bottom, so it's very portable. Setup is super easy - only one hose to the window, much less fiddly than those two-hose designs where you attach one hose and the other pops off.

Sure, the air conditioner has a major problem, but *it's not a major problem which most people will notice*. They may notice that most of the house is still hot, but the space

right in front of the air conditioner will be cool, so obviously the air conditioner is doing its job. Very few people will realize that the air conditioner is drawing hot air into the rest of the house. (Indeed, I saw zero reviews which mentioned that the air conditioner pulls hot air into the house - even the 1-star reviewers apparently did not realize *why* the air conditioner was so bad.)

[EDIT: several commenters seem to think that I'm claiming this air conditioner does not work *at all*, so I want to clarify that it will still cool down a room on net. If the air inside is all perfectly mixed together, it will still end up cooler with the air conditioner than without. The point is not that it doesn't work at all. The point is that it's stupidly inefficient in a way which I do not think consumers would plausibly choose over the relatively-low cost of a second hose if they recognized the problems.]

Generalization

Major problems are only fixed when those problems are obvious. Problems which most people won't notice (or won't attribute correctly) tend to stick around. There's no economic incentive to fix them.

And in practice, **there are plenty of problems which most people won't notice**. A few more examples:

- Most charities have pretty mediocre impact. But the actual impact is very-not-visible to the person making donations, so people keep donating. (Also people care about things besides impact, but nonetheless I doubt low-impact charities would survive if their ineffectiveness were generally obvious.)
- Medical research [has a replication rate below 50%](#). But when the effect sizes are expected to be small anyways, it's hard to tell whether it's working, so doctors (and patients) keep using crap treatments.
- Based on my firsthand experience with the B2B software industry, success is mostly determined by how good the product looks to managers making the decision to purchase. Successful B2B software (think "enterprise software") is usually crap, but has great salespeople and great dashboards for the managers.

... and presumably this extends to lots of other industries which I'm less familiar with.

Two points to highlight here:

- Regulation does not fix the problem, just moves it from the consumer to the regulator. A regulator will only regulate a problem which is obvious to the regulator. A regulator may sometimes have more expertise than a layperson, but even that requires that the politicians ultimately appointing people can [distinguish real from fake expertise, which is hard in general](#).
- Waiting longer does not fix the problem. All those people who did not notice their air conditioner pulling hot air into the house will not start noticing if we just wait a few years. Problems do not automatically become obvious over time.

How Does This Relate To Takeoff Speeds?

There's a common view that, as long as AI does not take off too quickly, we'll have time to see what goes wrong and iterate on it. It's a view with a lot of intuitive outside-view appeal: AI will work just like other industries. We try stuff, see what goes wrong, fix it. It worked like that in all the other industries, presumably it will work like that in AI too.

The point of the air conditioner is that other industries do not, in fact, work like that. Other industries are absolutely packed with major problems which are not fixed because they're not obvious. Even assuming that AI does not take off quickly (itself a dubious assumption at best), we should expect the same to be true of AI.

... But Won't Big Problems Be Obvious?

Most industries have major problems which aren't fixed because they're not obvious. But these problems can only be so bad. If they were really disastrous, the disasters would be obvious. Why not expect the same from AI?

Because AI will eventually be far more capable than human industries. It will, by default, optimize way harder than human industries are capable of optimizing.

What does it look like, when the optimization power is turned up to 11 on something like the air conditioner problem? Well, it *looks* really good. But all the resources are spent on looking good, not on actually being good. It's "[Potemkin village world](#)": a world designed to look amazing, but with nothing behind the facade. Maybe not even any living humans behind the facade - after all, even generally-happy real humans will inevitably sometimes appear less-than-maximally "good".

... But Isn't Solving The Obvious Problems Still Valuable?

The nonobvious problems are the whole reason why AI alignment is hard in the first place.

Think about the "game tree" of alignment - the basic starting points, how they fail, what strategies address the failures, how those fail, etc. The most basic starting points are generally of the form "collect data from humans on which things are good/bad, then train something to do good stuff and avoid bad stuff". Assuming such a strategy could be implemented efficiently, why would it fail? Well:

- In cases where humans label bad things as "good", the trained system will also be selected to label bad things as "good". In other words, the trained AI will optimize for things which look "good" to humans, even when those things are not very good.
- The trained system will likely end up implementing strategies which do "good"-labeled things in the training environment, but those strategies will not necessarily continue to do the things humans would consider "good" in other environments.

(Somewhat more detail on these failure modes [here](#).) Optimizing for things which look "good" to humans obviously raises exactly the sort of failure which the air conditioner points to. Failure of systems to generalize in "good" ways is less centrally about obviousness, but note that if it were obvious that the system were going to generalize badly, this would also be a pretty easy issue to solve: just don't deploy the system if it will generalize badly. Problem is, we can't tell whether a system will do what we want in deployment just by looking at what it does in training; we can't tell by looking at the system's behavior whether there's problems in there.

Point is: problems which are highly visible to humans are already easy, from an alignment perspective. They will probably be solved by default. There's not much *marginal* value in dealing with them. The value is in dealing with the problems which are hard to recognize.

Corollary: alignment is not importantly easier in slow-takeoff worlds, at least not due to the ability to iterate. The hard parts of the alignment problem are the parts where it's nonobvious that something is wrong. That's true regardless of how fast takeoff speeds are. And the ability to iterate does not make that hard part easier. Iteration mainly helps on the parts of the problem which were already easy anyway.

So I don't really care about takeoff speeds. The technical problems are basically similar either way.

... though admittedly I did not actually learn everything I need to know about takeoff speeds *just* from air conditioner ratings on Amazon. It took a lot of examples in different industries. Fortunately, there was no shortage of examples to hammer the idea into my head.

Mental Health and the Alignment Problem: A Compilation of Resources

This is a post about mental health and disposition in relation to the alignment problem. It compiles a number of resources that address how to maintain wellbeing and direction when confronted with existential risk. The target audience is anyone concerned with alignment, but the aim is for it to be particularly helpful for those who are new to these ideas. My sense is that most people who have known about the alignment problem for some time have already found their own strategies for resilience.

Many people in this community have posted [similar](#) ideas after Yudkowsky's "[Death With Dignity](#)" generated so much conversation on the subject. This post intends to be more touchy-feely, dealing more directly with emotional landscapes than questions of timelines or probabilities of success.

The resources section would benefit from community additions. Please suggest any resources that you would like to see added to this post and I will update it.

Please note that this document is not intended to supplement professional medical or psychological help in any way. Many preexisting mental health conditions can be exacerbated by these conversations. If you are concerned that you may be experiencing a mental health crisis, please [consult a professional](#).

Intro

There is no right way to emotionally respond to the reality of approaching superintelligent AI, and our collective responsibility to align it with our values.

There are days when I find it easy to be inspired by the difficulty of the alignment problem, and days where my life feels richer, more meaningful, and more charged for knowing about it.

There are other days when I find the whole affair terrifying and bleak, and days where I feel claustrophobic, sentenced to an impossibly cruel fate.

There are other days when I'm not in touch with the magnitude of the stakes, and where I am capable of acknowledging the alignment problem but have lost contact with it in my body and emotions.

I try not to pass judgment on how my sensitivity or vulnerability shifts day to day. But I recognize that my disposition matters. It has consequences on my motivation and actions, and frames how I perceive and interpret the world. And I recognize that learning about alignment has created a fundamental shift in my disposition; I experience the world in a different way now. The emotions I need to process (duty? zeal? terror? awe?) are, at the very least, more complicated.

I imagine it is similar for many of you, and will be similar for many more as they learn about the alignment problem and [the fact that we might not succeed^{\[1\]}](#). As that happens, I want to ensure we have the tools and resources to **be okay**. Here, the

valence of “be ok” is your decision. This question could be rephrased “how can I *thrive despite* the alignment problem,” “how can I *cope with* the alignment problem,” “how can I *overcome my fear of* the alignment problem,” etc. Everyone needs to [find their own question and their own answer](#).

At its foundation “being ok” is the decision to continue to live facing reality and the alignment problem directly, with internal stability and rationality intact. And as a high ideal, we’re going for some degree of inviolability, of unconditional wellbeing. The kind of wellbeing that holds onto “okayness” even if the possibility of solving alignment drops to 0.

As we learn to do this, we also gain the ability to be more helpful and compassionate to others. I have at times found myself distressed at someone’s inability to grasp the alignment problem the way I do, or at other times feeling almost a schadenfreude “I told you so,” when someone I previously talked to about alignment says “now I get it, and I’m scared.” We can do more good for the world and each other when we come from a place of compassion, understanding the sensitivity of each new mind that recognizes the alignment problem for what it is. Our ability to stand in some place of positive mental health and stability while facing the alignment problem directly can be difficult. It is a gift if we can do that for ourselves, and a gift if we can share it with others.

Fortunately we don’t have to do this alone, and this community is a good place to start. Many community members have found ways to make sense of themselves, their work, and their lives in relation to the alignment problem, and written about it in posts that we can draw inspiration from.

Resources

Below, I’ve put together many of the resources that I could find on the subject, summarized them, and broken them into categories for others to easily access. I’ve tried to pull together resources that have something specific to say about mental health under challenging circumstances like alignment.

That said, it was difficult to find posts directly relevant to this conversation as the majority of alignment-related content (and this is a good thing) focuses on the object-level rather than derivative issues like mental health.

If you know of other resources that would be useful to add to this list, or would like to correct any mischaracterization of resources I’ve currently listed, that would be helpful. Also, if you would have your own testimony/emotional approach to alignment, I invite you to share that in the comments below.

Text in *italics* below are quotations from the posts.

Alignment Positions

These posts explain orientations towards alignment that can help frame many aspects of our lives, from our time allocation and decision-making to how we choose to react to certain emotions that arise.

- **Ruby:** [A Quick Guide to Confronting Doom](#). Start here. This post is exactly about this subject, and is a good preface to reading the opinions below with the appropriate epistemic distance.

It's easy to hear the proclamations of [highly respected] others and/or everyone else reacting and then reflexively update to "aaahhhh". I'm not saying "aaahhhh" isn't the right reaction, but I think for any given person it should come after a deliberate step of processing arguments and evidence to figure out your own anticipations...My guess is that people who are concluding P(Doom) is high will each need to figure out how to live with it for themselves. My caution is just that whatever strategy you figure out should keep you in touch with reality (or your best estimate of it), even if it's uncomfortable.

- **Eliezer Yudkowsky:** [MIRI announces new 'Death with Dignity' strategy](#). Partially in jest, this post advocates that a good orientation for dealing with the alignment problem is to take actions that generate "dignity." There has been debate in the community over both the aesthetics and content of the post, but there's a coherent takeaway that I think most can agree on: try to be rational, and failing that, develop a good deontological strategy that protects against irresponsible action. There are a number of other Yudkowsky posts below which add nuance to this framework, and his [Coming of Age](#) sequence discusses alignment in [Beyond the Reach of God](#).

So don't get your heart set on that "not die at all" business. Don't invest all your emotion in a reward you probably won't get. Focus on dying with dignity - that is something you can actually obtain, even in this situation... the measuring units of dignity are over humanity's log odds of survival - the graph on which the logistic success curve is a straight line. A project that doubles humanity's chance of survival from 0% to 0% is helping humanity die with one additional information-theoretic bit of dignity.

The comments section of this post offers a number of meaningfully different perspectives, including AI_WAIFU's "[Fuck. That. Noise.](#)" disagreement.

- **Alex Turner:** [Emotionally Confronting a Probably Doomed World](#). A response to Yudkowsky's post, Turner argues that we should decouple our emotional response from the probability of doom, and escape the idea that we are "living in a tragedy." Turner's [Swimming Upstream](#) talks about earlier decisions to confront the alignment problem.

We do not live in a story. We can, in fact, just assess the situation, and then do what makes the most sense, what makes us strongest and happiest. The expected future of the universe is—by assumption—sad and horrible, and yet where is the ideal-agency theorem which says I must be downtrodden and glum about it?

- **Landfish:** [Don't die with dignity, instead play your outs](#). A response to Yudkowsky's post, Landfish argues for an MTG-inspired strategy of "playing your outs," or responding to a low-odds-of-success-future by looking ahead for what opportunities and affordances might still be available. Note that there is [disagreement](#) about the risk of this strategy and if it is consequentially useful for alignment; I include it here because it also offers an emotional frame.

I find playing to your outs a lot more motivating. The framing doesn't shy away

from the fact that winning is unlikely. But the action is “playing” rather than “dying”. And the goal is “outs” rather than “dignity”. Again, I think the difference is in connotation and not actually strategy.

- **Nate Soares:** [On Caring](#). Nate's post describes some of his motivation to be an effective altruist and do as much good as possible. While not specific to alignment, Nate works at MIRI and this view could be understood through that context as an argument for facing the alignment problem “courageously.”

Humanity is playing for unimaginably high stakes. At the very least, there are billions of people suffering today. At the worst, there are quadrillions (or more) potential humans, transhumans, or posthumans whose existence depends upon what we do here and now... Courage isn't about being fearless, it's about being able to do the right thing even if you're afraid. And similarly, addressing the major problems of our time isn't about feeling a strong compulsion to do so. It's about doing it anyway, even when internal compulsion utterly fails to capture the scope of the problems we face.

- **John Wentworth:** [WE CHOOSE TO ALIGN AI](#), and [The Plan](#). Together, these posts are a summary of Wentworth's emotional position with respect to alignment and his specific plan to work on the problem. He offers a perspective that the magnitude of the challenge is reason for inspiration, not despair.

When people first seriously think about alignment, a majority freak out. Existential threats are terrifying. And when people first seriously look at their own capabilities, or the capabilities of the world, to deal with the problem, a majority despair... but for someone who wants the challenge, the emotional response is different. The problem is terrifying? Our current capabilities seem woefully inadequate? Good; this problem is worthy. The part of me which looks at a rickety ladder 30 feet down into a dark tunnel and says “let's go!” wants this. The part of me which looks at a cliff face with no clear path up and cracks its knuckles wants this. The part of me which looks at a problem with no clear solution and smiles wants this. The response isn't tears, it's “let's fucking do this”.

- **Alex Flint:** [Musings on General Systems Alignment](#). Flint advocates that we envision the world as “fundamentally friendly to [AI alignment] efforts,” a perspective that allows us more optimism, resolve, and potentially more success in speaking with others about the issue.

It is not that our civilization has woken up completely to the dangers of advanced AI. It is that our civilization has not woken up, yet wishes to wake up, and knows that it wishes to wake up, and has found just enough clarity to bestow significant power and resources to us in the hope that we will take up leadership... Our job is to find the resolve to move forward with this difficult task, without getting caught up in the harmful patterns that exist in the world, and without losing track of the subtle way in which everyone is on our side.

- **Richard Ngo:** [My Attitude Towards Death](#). Ngo's post discusses fear of death, and his optimism for the future. He implies a strategy of “conversing” with his fear and trying to reassure it as a method for better integrating concerns.

What would happen if I talked more to the part [of me] that's scared of death, to try and figure out where it's coming from? By default, I expect it'd be

uncooperative - it wants to continue being scared of death, to make sure that I act appropriately (e.g. that I stay ambitious). Can I assure it that I'll still try hard to avoid death if it becomes less scared? One source of assurance is if I'm very excited about a very long life - which I am, because [the future could be amazing](#). Another comes from the altruistic part of me, whose primary focus is increasing the probability that the future will in fact be amazing. Since I believe that we face significant existential risk this century, working to make humanity's future go well overlaps heavily with working to make my own future go well. I think this broad argument (along with being in communities which reward longtermist altruism) has helped make the part of me that's scared of death more quiescent.

- **Holden Karnofsky:** [Call to Vigilance](#). In the last post in Karnofsky's [The Most Important Century](#) sequence, he describes that his emotional response to the alignment problem (and other challenges of our time) is one of intense, mixed emotions. He warns against people acting recklessly just to "do something," and advocates that people should remain aware and put themselves in positions to take robustly good actions.

When confronting the "most important century" hypothesis, my attitude doesn't match the familiar ones of "excitement and motion" or "fear and avoidance." Instead, I feel an odd mix of intensity, urgency, confusion and hesitance. I'm looking at something bigger than I ever expected to confront, feeling underqualified and ignorant about what to do next. This is a hard mood to share and spread, but I'm trying.

- Please suggest additional "positions" and I will add them here.

Mental Strategies

These posts provide guidance for stabilizing positive states, and for transforming negative emotions that may arise when facing existential risks such as anxiety, fear, despair, apathy, and depression.

- **On rising to the challenge:** Yudkowsky and others: [Challenging the Difficult](#) and [Heroic Responsibility](#). As demonstrated by the "positions" above, many people find that the best way to counter negative emotions around the alignment problem is to work on it directly. These resources advocate for building the internal drive to tackle problems as serious as alignment, and may be especially useful for individuals struggling with "helplessness" by transforming that feeling into action.

"You could call it heroic responsibility, maybe," Harry Potter said. "Not like the usual sort. It means that whatever happens, no matter what, it's always your fault. Even if you tell Professor McGonagall, she's not responsible for what happens, you are. Following the school rules isn't an excuse, someone else being in charge isn't an excuse, even trying your best isn't an excuse. There just aren't any excuses, you've got to get the job done no matter what." -HPMOR, chapter 75.

If you're motivated to do something about alignment, there are [many pragmatic posts](#) on LW as well as non-LW resources like [AI Safety](#).

[Support](#), the [AGI Safety Fundamentals Course](#), and [80,000 Hours](#).

- **On how to overcome negative emotions:** Nate Soares: [Replacing Guilt](#). This foundational sequence has helped many people in the LW community transform feelings of guilt, resistance, sorrow, imposter syndrome, and other negative emotions into inspiration. Despite the title, its scope is much larger than “guilt” and is a great starting place for any reader. One post to call out is [Detach the Grim-o-meter](#), which offers the perspective that one can continue to be “curious, playful, and relaxed” even when looking directly at existential risks.

When all is said and done, Nature will not judge us by our actions; we will be measured only by what actually happens. Our goal, in the end, is to ensure that the timeless history of our universe is one that is filled with whatever it is we're fighting for. For me, at least, this is the underlying driver that takes the place of guilt: Once we have learned our lessons from the past, there is no reason to wrack ourselves with guilt. All we need to do, in any given moment, is look upon the actions available to us, consider, and take whichever one seems most likely to lead to a future full of light.

- **On accepting sorrow and fear:** Yudkowsky's [Feeling Rational](#) and Luke Muehlhauser's [Musks' Non-missing Mood](#). In contrast with some of the posts above that encourage decoupling emotions from probabilities of doom, these posts offer the perspective that negative emotions are not only a natural, but also a rational response. For those confronting negative emotions who would rather accept and work with those feelings than transform them, these posts may offer some insight.

When something terrible happens, I do not flee my sadness by searching for fake consolations and false silver linings. I visualize the past and future of humankind, the tens of billions of deaths over our history, the misery and fear, the search for answers, the trembling hands reaching upward out of so much blood, what we could become someday when we make the stars our cities, all that darkness and all that light—I know that I can never truly understand it, and I haven't the words to say. Despite all my philosophy I am still embarrassed to confess strong emotions, and you're probably uncomfortable hearing them. But I know, now, that it is rational to feel. - Feeling Rational

- **On working with imposter syndrome:** Seven years ago, Luke Muehlhauser recommended [If you're an “AI safety lurker,” now would be a good time to de-lurk](#). But imposter syndrome and self-doubt can prevent people from raising their hand. Yudkowsky's [Hero Licensing](#) talks about his own experience questioning the value of his work. Scott Alexander's [Parable of the Talents](#) and Nicole Ross' [Desperation Hamster Wheels](#) are not specific to alignment, but offer some advice on how to work with feelings of inadequacy. That said, self-worth is a deeper subject than just imposter syndrome and likely needs to be addressed outside of the context of productivity entirely, not to mention alignment.

When someone feels sad because they can't be a great scientist, it is nice to be able to point out all of their intellectual strengths and tell them “Yes you can, if only you put your mind to it!” But this is often not true. At that point you have to say “f@#k it” and tell them to stop tying their self-worth to being a great scientist. And we had better establish that now, before transhumanists succeed in creating superintelligence and we all have to come to terms with our

intellectual inferiority. - Parable of the Talents

- **On being honest about concerns:** Katja Grace [Beyond fire alarms: freeing the groupstuck](#). This post is primarily a response to Yudkowsky's "[There is No Fire Alarm for AGI](#)," but offers relevant ideas for how to deal with situations where one is afraid of looking silly for being overly-concerned about AI risk.

Practice voicing your somewhat embarrassing concerns, to make it easier for others to follow (and easier for you to do it again in future)... React to others' concerns that don't sound right to you with kindness and curiosity instead of laughter. Be especially nice about concerns about risks in particular, to counterbalance the special potential for shame there. [or about people raising points that you think could possibly be embarrassing for them to raise]

- **On overcoming avoidance:** Anna Salaman's [Flinching away from the Truth](#) and [Making your explicit reasoning trustworthy](#). Sometimes people avoid thinking about the alignment problem due to the concern of mistaken beliefs, going down lines of thinking that lead to seductive but inaccurate conclusions. Anna's posts may offer reassurance for those who are hesitant to engage fully to trust their own reasoning and not rely on others' positions.

"I don't want to think about that! I might be left with mistaken beliefs!" tl;dr: Many of us hesitate to trust explicit reasoning because we haven't built the skills that make such reasoning trustworthy. Some simple strategies can help.

- **On facing death:** In addition to Ngo's [My Attitude Towards Death](#), there are a number of LW posts on [death](#) that may be useful to this conversation. Some such as Joe Carlsmith's [Thoughts on Being Mortal](#) aren't about alignment but confront fear of death directly, while some such as Yudkowsky's [The Meaning that Immortality Gives to Life](#) touch on the singularity but are more about avoiding death. Avoidance of death is likely the crux of most fear and sorrow around the alignment problem, so from a purely mental-health related standpoint, it may be meaningful to try to separate emotional response to death from the alignment problem itself. Finding ways to confront death directly may afford a deep inviolability to existential fear.

*Sometimes, on the comparatively rare occasions when I experience even-somewhat-intense sickness or pain, I... am brought more directly into the huge number of subjective worlds filled with relentless, inescapable pain. These glimpses often feel like a sudden shaking off of a certain kind of fuzziness; a clarifying of something central to what's really going on in the world; and it also comes with fear of just how helpless we can become. - *Thoughts on Being Mortal**

- **HPMOR advice on facing existential risk:** Yudkowsky's [Harry Potter and the Methods of Rationality](#) is not about AI alignment (Harry deals mostly with local-to-planetary-scale rather than cosmological/hyperexistential threats), but the depicted emotions and mental strategies have [direct analogue](#). The story contains deep explorations of the internal experience of facing seemingly impossible odds, the burden of heroic responsibility, difficult tradeoffs and the necessity of sacrifice, and the motivation for rational action and self-improvement. A non-exhaustive list of chapters that might be useful (and would be much more useful in context with the whole sequence):

[Ch 39](#): death, motivation for transhumanism

- [Ch 43-46](#): fear, death, motivation for transhumanism
- [Ch 56-58](#): optimizing against improbable odds, despair
- [Ch 63](#): the burden of responsibility, longing for a normal life
- [Ch 75](#): heroic responsibility
- [Ch 79-82](#): sacrifice
- [Ch 88](#): fear of expressing panic, bystander apathy
- [Ch 89](#): accepting/rejecting an unacceptable reality
- [Ch 110](#): guilt, shame
- [Ch 111-115](#): optimizing against improbable odds, despair
- [Ch 117](#): guilt, sacrifice

- **EA resources on general wellbeing and burnout.** While not specific to alignment, it would be a mistake not to mention the wealth of information on the EA forum related to mental health such as Miranda Zhang's [Mental Health Resources tailored for EAs \(WIP\)](#) and Ewelina Tur's [List of Mental Health Resources](#). The EA forum also has a bunch of specific posts on burnout, like Elizabeth's [Burnout](#), Tessa's [Aiming for the minimum of self-care is dangerous](#), and Julia Wise's [Cheerfully](#).

Tools and Practices

These practices span from interventionist practices aimed at quickly cutting through negative states to longer-term practices aimed at building up more sustainable wellbeing.

- **Meditation:** Kaj Sotala's [My attempt to explain Looking, insight meditation, and enlightenment in non-mysterious terms](#). This post explains how meditation practices can help people develop unconditional equanimity, even in the face of existential risk. Sotala's [sequence](#) goes deeper in these ideas, and he also has a number of practical individual posts like [Overcoming suffering: emotional acceptance](#).

But if you cared about things like saving the world, then you will still continue to work on saving the world, and you will be Looking at things which will help you save the world - including ones that increase your rationality. It's just that if the world ends up ending, it won't feel like the end of the world. Of course, you will still feel intense grief and disappointment and everything that you'd expect to feel about the world ending. Intense grief and disappointment just won't be the end of the world. - My attempt to explain Looking

- **Productivity Sprints:** Nate Soares' [The mechanics of my recent productivity](#) and Logan Riggs' [Saving the world in 80 days](#) and [epilogue](#). Given that many of the above posts advocate for getting involved as a way to respond to the magnitude of the alignment problem, it's useful to have testimonials about how people got started. Nate's post provides some practical advice, and Logan's demonstrates what it looks like to put that advice into practice. Alex Turner's [Problem relaxation as a tactic](#) is also helpful for those looking to get into alignment who find the scope of the problem too large.

A decade ago, I decided to save the world. I was fourteen, and the world certainly wasn't going to save itself. I fumbled around for nine years; it's

surprising how long one can fumble around. I somehow managed to miss the whole idea of existential risk and the whole concept of an intelligence explosion... A year ago, I finally read the LessWrong sequences... On Saturday I was invited to become a MIRI research associate. It's been an exciting year, to say the least. - The mechanics of my recent productivity

- **Focusing and Noticing:** If you're not sure what you "feel" about the alignment problem, or emotions you've previously felt are out of reach, the techniques of [Focusing](#) and [Noticing](#) can help. These methods bring awareness to sensations within the body, which increases clarity and affords an opportunity to do something about the feelings. For example, it may help uproot unconscious motivations that may be driving undesirable habits (procrastination, doomscrolling, etc.), or may help with [anxiety or self-doubt](#).

Focusing refers to a family of introspective techniques... whose aim is to access one's "gut" or "System 1" feelings. Archetypically, sensations within the body are approached with a spirit of gentle curiosity, and possible verbal labels are checked against felt senses. Where successful, this can improve internal understanding and allow split off trauma or conflict between subagents to be processed for improved internal alignment.

- **Dark Arts:** Nate Soares' [The Dark Arts of Rationality](#). "Dark Arts" is a term for methods which involve deception or believing untrue things, such as intentional compartmentalization, inconsistency, or modifying terminal goals. These are not methods I'd recommend for everyone, but may help some individuals balance their life/productivity with intense feelings related to the alignment problem.

We are fortunate, as humans, to be skilled at compartmentalization: this helps us work around our mental handicaps without sacrificing epistemic rationality. Of course, we'd rather not have the mental handicaps in the first place: but you have to work with what you're given. We are weird agents without full control of our own minds. We lack direct control over important aspects of ourselves. For that reason, it's often necessary to take actions that may seem contradictory, crazy, or downright irrational.

People Resources

Unfortunately, this section is quite bare at the moment. I am interested in using it to promote therapists, coaches, instructors, or other individuals who provide support to those who may be struggling with their reactions to the alignment problem. Please let me know if you or someone you know would be a good fit for this list.

- [Shay Gestal](#) is a trained health coach who partners with [AI Safety Support](#) to offer completely free sessions to people working on AI safety.
- [EA Mental Health Navigator](#) has a list of coaches and therapists who have experience working with effective altruists. [Lynette Bye](#) is one of them, and has written a post with [advice for how to select a therapist](#). [Tee Barnett](#) is another coach with experience discussing existential risk.
- More TBD

A Final Note

One can make a case that robust mental health is instrumental for working on alignment. And that's certainly true. If you're reading this, it's probably because you are in a community that is working on alignment or adjacent to it. And if you're not, but are interested in finding more ways to help, this post points to resources to get started. One of the benefits of robust mental health practices is that it creates the stability necessary to continue to do the necessary work.

But this post is written with the intention of increasing wellbeing, not productivity. We work on the alignment problem because we are driven by our deep care to protect the world we know, the one in which people are capable of joy and beauty and love. I believe it's meaningful for us to abide in wellbeing and flourish for no other reason than that it is what life is about. Wellbeing is instrumental for solving alignment, but more importantly, wellbeing is why we're trying to solve it.

1. ^

While there is disagreement about the timeline on which we need to solve this problem, and disagreement about our probability of success, there's broad acceptance in this community that the problem is real and the stakes dire. Even those who are optimistic about things going well could imagine worlds in which they don't.

Emotionally Confronting a Probably-Doomed World: Against Motivation Via Dignity Points

This article was written [in ignorance](#) of the alignment community's reaction to Eliezer's "[Death with Dignity](#)" post. The first part of this article responds to how I suspect some people reacted to that post, while the second part is my take on the post itself.

I write against defeatism; I write against decline; I write against that internal slumping that sneaks in on the coat-tails of bad news. I do not dispute the balance of evidence—*are we doomed, or not?* Let us simply assume that we live in a relatively doomed world: It's very improbable that we solve AI alignment in time.

What next?

We have been taught what comes next, in this kind of story. Since, by assumption, we won't receive a "happily ever after", we infer we are in a *tragedy*. Realizing this, we are disappointed and sad. The fellowship breaks and scatters, its once-proud and vibrant members downtrodden. And then occurs a miracle which we could have turned to our advantage, but for our civilizational incompetence and our own stupidity—smart enough to build AI, yet too dumb to align it. And then the laugh track plays. And then we die! The end!

As AI_WAIFU [said](#): "Fuck. That. Noise."

We do not live in a story. We can, in fact, just assess the situation, and then do what makes the most sense, what makes us strongest and happiest. The expected future of the universe is—by assumption—*sad and horrible*, and yet where is the ideal-agency theorem which says I must be downtrodden and glum about it?

Against retreat

In response to how I suspect some people reacted to "Death with Dignity."

Suppose we have ten years left. Some may consider a very natural response—to retreat. Spend the remaining time with your family and your friends, working whatever job you want—just do whatever.

But... Um... This sounds like an awful plan? I'd feel like a lazy bum, like a cornered animal. And for those of us doing direct work, this plan sounds *stupid*. It sounds like a great way to *throw away worlds where we would receive a miracle, and yet couldn't do anything about it anymore because we all went home*.

How would I like to see those ten years spent? I'm looking for plans which involve *playing to win*, which are *fun* and *edifying*, which would make me *proud* to be me and to be part of this community.

Ladies and gentlemen, let me mention a concept called *instrumental convergence*. [It's something I like thinking about sometimes](#). In many situations, for many goals, the best actions look similar: Gather resources, build up strength, develop yourself. Following this pattern, good plans in probably-doomed worlds look a lot like good plans in hopeful worlds.

Communal edification

Even if the probable-end is bearing down—imagine having a vibrant social scene *anyways*, where people are having fun *anyways*, developing deep friendships *anyways*, expecting more of themselves *anyways*, pushing each other to break their boundaries *anyways*. Because that's awesome, because that's fun, because that makes us stronger, because that positions us to seize miracles, because that's the kind of thing that makes me proud, dammit.

Our civilization is incompetent. Gee, that sucks. But we don't have to suck too, do we? If a society of people-like-us would really be so great, [why don't we show it?](#)

- Set up communal institutions which *actually have good incentives*.
- Build out infrastructure for alignment research and training, so that smart young people want to try their hand at alignment research.
 - Probably more people should skill up as managers, so we can have more capacity via high-quality organizations like Redwood Research. I don't know how to do this, and cannot offer advice.
- Set up infrastructure in areas where our society is decrepit—can we do better
 - on childcare?
 - on healthcare? (amortize the fixed costs of medical tourism over a pool of individuals?)
- Build stronger and better and smarter norms
 - around event culture,
 - around dating
 - I applaud Eliezer for funding a community matchmaker last year; I want more people trying ideas!
- Brainstorm over and over and over what institutions may help with miracles, and then build them. Offer large cash prizes for ideas which top thinkers agree is valuable.

Of course, we also want an AI alignment training pipeline full of brilliant bright-eyed newcomers, with obvious improvements to the pipeline rewarded by cash prizes (or a prediction market-based scheme). We list the possible miracles by probability, and find the prerequisites to seizing them, and then train a distribution of researchers according to these expected prerequisites. We train an army of alignment researchers—even if we don't see any good research pathways now, because that can change later.

And yet it seems to me like Eliezer himself does not encourage some important preparations. In [this comment](#) and in other hearsay I am privy to, he has little advice to offer prospective alignment researchers, which I think is because he is unsure what will help.

Well, what he says about alignment is his business. But it's also *my* business, since I live on this planet and share its fate. And even if I agreed completely with Eliezer, I think I would rather give advice like:

I think the situation is very bleak, and assess that you individually have a low probability of saving the world. I think you should know that before starting. That said, if you are still interested in helping, if you have something to protect—you should master deep learning and the science of intelligence, and you should practice the art of rationality with all the intensity and honesty you can muster.

If you want to do alignment theory, try the AI safety camp. Or maybe see if Chris Olah could use some help. And perhaps, one day, we will have a worthy alignment scheme which urgently needs the expertise which you can develop today.

As I see things, this advice is both honest and pragmatic—it is useful in many “miracle” worlds. Wouldn’t it be undignified to later have a massive breakthrough in alignment theory, and yet not have enough hands on deck to derive enough corollaries and evaluate alignment schemes derived thereby?

As for the individual—

Individual edification

1. If you are having a hard time right now—that is OK; please do not use your anxiety/guilt as a cudgel to beat yourself into yet deeper anxiety/guilt. Behind these posts is a person who also sometimes has a hard time.
2. If you feel that you *should* do more to help—not due to a deep yearning, but because you’d be *bad* if you didn’t do more—please read [Replacing Guilt](#). I do not want bedraggled people pushing themselves *yet harder*, working *even more hours*, in order to scrap out a win. That’s unhealthy, and not an awesome or fun way to treat yourself.^[1]
3. It’s hard to advise other people. I’m going to talk about what I want from myself. Translate accordingly.

Along those lines, I want to become as strong as I can. Maybe it won’t be enough to win, but maybe it will. And hot damn, do I want to become strong. Because it’s awesome, because it’s fun, and because I want to win.

Anyways, what do I mean by “strong”? Do I mean that I’ve absorbed lots of textbooks, that I know lots of facts, that I have lots of impressive skills? No. Here, “strong” means strong at *the basics of thinking*:

Everything inside Keltham’s mind has a very trained feeling to it, his moment-to-moment thought-motions each feeling like a punch that a monk throws after twelve years of experience in martial arts, when the monk isn’t particularly focused on showing off and simply knows what he’s doing without thinking about it.

Mad Investor Chaos

I think that many people think that becoming more rational involves being *harder on yourself*; imagining yet fiercer and nitpickier critics, and coming out unscathed anyways because you were so paranoid. You checked more boxes, defended against more biases, and examined the evidence even harder.

That's... not how it works. I may write elsewhere about how I have found it to work. But I will say this: More is possible.

I used to think I had absorbed the Sequences, that I had acquired most of the available art of rationality, and the rest I would have to build or experience for myself. I was wrong. Early this year, my brain got pinged [in just the right way](#) by Eliezer's *Mad Investor Chaos* fiction, and—click!

- Before, I had declarative knowledge of biases and knew how they felt from the inside; I knew the basic math of probability theory; I had experience resolving internal motivational conflicts.
- After, I had an intuitive sense for (extremely basic) probability theory, a bright standard in my mind against which I compare my thoughts. “Oh, I believe *that* quite strongly. Why? Where did I get the evidence-fuel for the strong previous-update likelihood ratio implied by this present credence?” These concepts feel *atomic* to me, even though they’re clumsy to write, and eg the odds form of Bayes’ rule feels like a primitive mental operation.

I could see the ways in which my mental footwork is sloppy. The sloppiness made me realize how unoptimized my thought processes are—I had never optimized that!—and how much more I could learn. And so I began, and now I have a small part of me which is increasingly consequentialist and Bayesian, a part which I can call upon for strength and clarity.

And this is what I have done in a few months. Who could I grow into in ten years? I want to become strong, and I want to search for plans which *win*. Because trying to die in a “dignified” way is not a wise strategy (for people like me, at least).

Against motivation via dignity points

When Earth's prospects are that far underwater in the basement of the logistic success curve, it may be hard to feel motivated about continuing to fight, since doubling our chances of survival will only take them from 0% to 0%.

That's why I would suggest reframing the problem - especially on an emotional level - to helping humanity die with dignity, or rather, since even this goal is realistically unattainable at this point, die with slightly more dignity than would otherwise be counterfactually obtained.

...

So don't get your heart set on that "not die at all" business. Don't invest all your emotion in a reward you probably won't get. Focus on dying with dignity - that is something you can actually obtain, even in this situation. After all, if you help humanity die with even one more dignity point, you yourself die with one hundred dignity points! Even if your species dies an incredibly undignified death, for you to have helped humanity go down with even slightly more of a real fight, is to die an extremely dignified death.

...

But if enough people can contribute enough bits of dignity like that, wouldn't that mean we didn't die at all? Yes, but again, don't get your hopes up. Don't focus

your emotions on a goal you're probably not going to obtain. Realistically, we find a handful of projects that contribute a few more bits of counterfactual dignity; get a bunch more not-specifically-expected bad news that makes the first-order object-level situation look even worse (where to second order, of course, the good Bayesians already knew that was how it would go); and then we all die.

- [MIRI announces new "Death With Dignity" strategy](#)

Again, I am not, in this post, disputing Eliezer's object-level model. I have supposed he is correct about our probable doom. Obviously, if we are in a probably-doomed world, I will keep that in mind. I do actually want to *win*, and finding winning plans requires entangling my brain with the details of each expected danger. Sharp danger awaits when you lose sight of the fact that reality is allowed to kill you.

However—if you work at all like I do, I think this is *not* how you should interface with yourself or your motivational system.^[2] It is like saying:

This vault door was professionally secured, and you are no professional burglar. Yes, your mother is starving inside, but you are unlikely to open the door before she dies. Therefore, you should search for *dignified* plans—plans which let you seize miracles in worlds where the door was installed incorrectly. But please don't expect to actually open the door. Don't get your hopes up.

A search for *dignified plans* is different from a search for plans which *get my mother out of the damn vault*. I can, in fact, conduct the latter search while *still remembering how unlikely I am to actually open the vault*, and the latter search has a *better chance of actually finding dignified plans!*

Want to try to make a million dollars? Buy a [lottery ticket](#). Your odds of winning may not be very good, but you did try, and trying was what you wanted. In fact, you tried your *best*, since you only had one dollar left after buying lunch.

Maximizing the odds of goal achievement using available resources: is this not intelligence?

It's only when you want, above all else, to *actually flip the switch*—without quotation and without consolation prizes just for trying—that you will *actually* put in the effort to *actually* maximize the probability.

But if all you want is to “maximize the probability of success using available resources,” then that’s the easiest thing in the world to convince yourself you’ve done. The very first plan you hit upon will serve quite well as “maximizing”—if necessary, you can generate an inferior alternative to prove its optimality. And any tiny resource that you care to put in will be what is “available.” Remember to congratulate yourself on putting in 100% of it!

Don’t try your best. Win, or fail. There is no best.

[Trying to Try](#) ^[3]

I hope you do not let this “dignity” orientation cloud your intent to win. In fact, I think you should lean in the opposite direction, and sharpen your intent. *Search only for plans which actually win, given your best understanding of the AI alignment problem. Be honest with yourself; do not flinch away from reality; do not take your eyes off the goal.*

I'm not going to waste my time searching for dignified plans (which maximize humanity's probability of survival). Because I do have a mother in that vault, and a father, and a brother. In fact, there's a whole damn planet in there. It's my home, and it's yours, too. And if we do stare down defeat together—let's make that remaining time valiant and exciting and awesome.

1. ^

Eliezer wrote:

We cannot fight at maximum all the time, and some times are more important than others. (Namely, when the logistic success curve seems relatively more sloped; those times are relatively more important.)

This is a good point. I am not advocating burnout. This is indeed a resource to conserve for situations closer to the 50% success rate, or for a targeted push at a particularly impactful moment. I am advocating growth and development in a way which is *fun and awesome*, which pushes limits without straining endurance. Perhaps this is not how most people work. But it's how I can work.

2. ^

Furthermore, do not [conflate feelings with beliefs](#). You do not have to believe the future is rosy in order to feel good and strong and healthy and to give the problem everything you've got (without pushing to burnout). Feelings are not beliefs! I think of certain feelings as (harder-to-control) *actions*.

3. ^

The point is not that Eliezer said “do not ‘try your best’” and now he has pivoted away from win-oriented-thinking, *aha, gotcha!*—The point is that I think the original orientation is wise and healthy and strengthening, and the new orientation is not.

Only Asking Real Questions

It's easy to accidentally ask children questions that aren't real questions. Perhaps the right answer is so obvious I don't realize in time that they might decide otherwise: "do you want to wait to eat your messy candy until we're out of the car?" Or I don't figure out that there was only one answer that would work until after I finish asking the question: "do you want to do your bath now, or before dinner? ... Oh, wait, we have guests coming over in the afternoon and we can't do the bath then, now's the only time that works."

I find that these interactions tend to go poorly, much worse than if I'd never asked a question. Something like, they thought they had a choice, try to make a choice, and now they're feeling me take that away.

I've tried pretty hard to get myself to stop doing this, and one strategy that has worked for me was deciding to stand by my mistakes. If I accidentally offer them a choice when I shouldn't have, I won't withdraw it if they choose (to my mind) badly. This isn't a completely strict rule: I can imagine foolish-enough things that I could accidentally ask that it wouldn't make sense to stand by. On the other hand, though, I've found I can recover from most things if I'm willing to put some extra effort. Going back to the messy candy example above, and imagining some further discussion:

- Parent: Do you want to wait to eat your messy candy until we're out of the car?
- Child: I want to eat it now!
- Parent: That's gonna make a mess everywhere. I don't think that's a good idea.
- Child: But I want the candy now!
- Parent: If it makes a mess, will you clean it up?
- Child: No.
- Parent: How will it get cleaned up?
- Child: You can do it.
- Parent: I don't want to clean it up.
- Child: I don't want to either.
- Parent: If you make a mess, it's your job to clean it up.
- Child: Ok, I'll do it.
- Parent: Will you put up a \$1 [bond](#)? Where I get to keep the dollar if you don't clean up?
- Child: Ok. [eats messy candy]
- [time passes, arrival, getting out of the car]
- Parent: It looks like there's a mess on the seat.

- Child: I don't want to clean it up.
- Parent: If you don't, I will keep your dollar.
- Child: [cleans up, slowly, with parental help depending on age]

This definitely requires a lot more time and back-and-forth than just saying "sorry, I shouldn't have offered that, we shouldn't eat messy candy in the car". But it avoids going back on what you offered, is more respectful, and also gives an opportunity to show why you didn't think it was a good idea.

Comment via: [facebook](#)

Supervise Process, not Outcomes

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

We can think about machine learning systems on a spectrum from process-based to outcome-based:

- Process-based systems are built on human-understandable task decompositions, with direct supervision of reasoning steps.
- Outcome-based systems are built on end-to-end optimization, with supervision of final results.

This post explains why [Ought](#) is devoted to process-based systems. The argument is:

1. In the short term, process-based ML systems have better differential capabilities: They help us apply ML to tasks where we don't have access to outcomes. These tasks include long-range forecasting, policy decisions, and theoretical research.
2. In the long term, process-based ML systems help avoid catastrophic outcomes from systems gaming outcome measures and are thus more aligned.
3. Both process- and outcome-based evaluation are attractors to varying degrees: Once an architecture is entrenched, it's hard to move away from it. This lock-in applies much more to outcome-based systems.
4. Whether the most powerful ML systems will primarily be process-based or outcome-based is up in the air.
5. So it's crucial to push toward process-based training now.

There are almost no new ideas here. We're reframing the well-known outer alignment difficulties for traditional deep learning architectures and contrasting them with compositional approaches. To the extent that there are new ideas, credit primarily goes to Paul Christiano and Jon Uesato.

We only describe our background worldview here. In [a follow-up post](#), we explain why we're building [Elicit](#), the AI research assistant.

The spectrum

Supervising outcomes

Supervision of outcomes is what most people think about when they think about machine learning. Local components are optimized based on an overall feedback signal:

- SGD optimizes weights in a neural net to reduce its training loss
- Neural architecture search optimizes architectures and hyperparameters to have low validation loss
- Policy gradient optimizes policy neural nets to choose actions that lead to high expected rewards

In each case, the system is optimized based on how well it's doing empirically.

[MuZero](#) is an example of a non-trivial outcome-based architecture. MuZero is a reinforcement learning algorithm that reaches expert-level performance at Go, Chess, and Shogi without human data, domain knowledge, or hard-coded rules. The architecture has three parts:

- A representation network, mapping observations to states
- A dynamics network, mapping state and action to future state, and
- A prediction network, mapping state to value and distribution over next actions.

Superficially, this looks like an architecture with independently meaningful components, including a “world model” (dynamics network). However, because the networks are optimized end-to-end to jointly maximize expected rewards and to be internally consistent, they need not capture interpretable dynamics or state. It’s just a few functions that, if chained together, are useful for predicting reward-maximizing actions.

Neural nets are always in the outcomes-based regime to some extent: In each layer and at each node, they use the matrices that make the neural net as a whole work well.

Supervising process

If you’re not optimizing based on how well something works empirically (outcomes), then the main way you can judge it is by looking at whether it’s structurally the right thing to do (process).

For many tasks, we understand what pieces of work we need to do and how to combine them. We trust the result because of this reasoning, not because we’ve observed final results for very similar tasks:

- Engineers and astronomers expect the James Webb Space Telescope to work because its deployment follows a well-understood plan, and it is built out of well-understood modules.
- Programmers expect their algorithms to implement the intended behavior because they reason about what each function and line does and how they go together to bring about the behavior they want.
- Archeologists expect their conclusions about the age of the first stone tools to be more or less correct because they can reason about the age of the sediment layer the tools are in. They can estimate the age of the layers by looking at the iron-bearing minerals they contain which reflect the state of the earth’s magnetic polarity at the time they were preserved.

At Ought, we’ve been thinking about scientific literature review as a task that we expect to arrive at correct answers only when it’s based on a good process. When I’m trying to figure out whether iron supplements will help me or hurt me, I might start by following a process like this:

1. Clarify the question I’m trying to answer—what kind of iron, what kinds of supplements, what benefits am I hoping for? How will I decide whether to take the supplement or not?
2. Search for a list of candidate papers using the question and related search terms
3. For each study I find, answer:

1. Does it address the question I'm interested in, or a closely related question? Was the population studied similar to me?
2. Is it a randomized controlled trial, or a meta-analysis of trials?
3. Is the risk of bias below the threshold I'd accept? Are there no glaring critiques of the study or methodological limitations?
4. Throw out studies for which the answer isn't yes to all questions
5. If any studies remain, synthesize them into a summary answer that explains the observed evidence
6. If not, relax my question and go back to 2

Of course, this is far from a great process. For a slightly better example, see [this systematization](#) of Scott Alexander's [post](#) on Indian Economic Reform.

To build a process-based system, the fundamental problem to solve is to reduce the long-horizon tasks we care about to independently meaningful short-horizon tasks ([factored cognition](#)). If we can do that, we can then generate human (or human-like) demonstrations and feedback for these sub-tasks.

This reduction to subtasks can be done by the system designer, or for better scalability on-the-fly by the system itself. Task decomposition is another subtask, after all.

In between process and outcomes

Many tasks can be approached in both ways, and in practice, most systems will likely end up somewhere in between. Examples:

Search engine:

- Outcome-based: Embed documents and metadata in a vector space, same for queries. Use a neural net retriever. Optimize retriever parameters and embeddings for users giving high ratings to the retrieved documents.
- Process-based: Define an idealized process for evaluating the quality of a search result for a given query, e.g. decomposing the evaluation of the question “Is this result trustworthy?” into Pagerank-style considerations, questions like “Is the author an expert in this field?” Distill each of the subquestions modules into a neural net so that we can execute it at runtime.
- In between: Start with the process-based approach, but use user scores to make a few choices, such as fitting parameters in a tiny MLP that mixes feature weights.

Question-answering:

- Outcome-based: Train a neural net to map questions to answers, perhaps using [Retro-style](#) end-to-end-optimized retrieval.
- Process-based: Independently train neural nets that map questions to web search queries, query responses to relevant extracts, and long answers to summary answers, each trained on human demonstrations or feedback.
- In between: Follow the process-based approach, but (as in [WebGPT](#)), don't imitate human queries; instead, just learn query strategies that lead to highly rated final answers end-to-end.

Business decision advisor:

- Outcome-based: Train a MuZero-style neural net on making decisions about trades, product launches, and hiring decisions based on business returns, or other long-term metrics of interest, optimizing for actions that look good in hindsight.
- Process-based: Imitate actions that look good to a human supervisor in foresight, giving the human supervisor AI tools to do a thorough ex-ante evaluation.
- In between: Imitate human actions chosen ex-ante, but use predicted long-term metrics to choose between several actions that all look similarly good in foresight.

Eric Drexler's [CAIS](#) paints a picture of AI that is also somewhere between process and outcomes in that AI services have clearly defined roles on a larger scale, but are individually outcome-based.

It's better to supervise process than outcomes

Why prefer supervision of process? If we don't need to look at outcomes, then:

1. We can do well at long-horizon tasks where outcomes aren't available (better differential capabilities)
2. We don't run the risk of our outcome measures being gamed (better alignment)

Differential capabilities: Supervising process helps with long-horizon tasks

We'd like to use AI to advance our collective capability at long-horizon tasks like:

- Multi-year and multi-decade forecasting, e.g., predicting long-term consequences of vaccines
- Policy and governance, especially AI policy
- Personal and institutional planning and decision-making
- AI alignment research

Unfortunately, gathering outcome data is somewhere between expensive and impossible for these tasks. It's much easier to gather data and exceed human capability at short-horizon tasks:

- Keeping people engaged as they interact with videos and posts
- Developing physical technologies, e.g., new toxic molecules
- Persuading people in short conversations
- Predicting 30-minute consequences, not 30-year consequences

In a world where AI capabilities scale rapidly, we need AI to support research and reasoning that is likely to make AI go better. This includes guiding AI development and policy, helping us figure out what's true and make plans as much as it helps us persuade and optimize goals with fast feedback loops and easy specifications.

If we can reliably reduce such long-horizon tasks to short-horizon tasks, we'll be better positioned to deal with the incremental development and deployment of advanced AI.

Alignment: Supervising process is safety by construction

With outcome-based systems, we'll eventually have AI that is incentivized to game the outcome evaluations. This could lead to catastrophes through AI takeover. (Perhaps obvious to most readers, but seems worth making explicit: A big reason we care about alignment is that we think that, from our current vantage point, the world could look pretty crazy^[1] in a few decades.)

What is the endgame for outcome-based systems? Because we can't specify long-term objectives like "don't cause side-effects we wouldn't like if we understood them", we're using proxy objectives that don't fully distinguish "things seem good" from "things are good". As ML systems get smarter, eventually all of the optimization effort in the world is aimed at causing high evaluations on these proxies. If it's easier to make evaluations high by compromising sensors, corrupting institutions, or taking any other bad actions, this will eventually happen.

Suppose instead that we understood the role of each component, and that each component was constructed based on arguments that it will fulfill that role well; or it was constructed and understood by something whose behavior we understood and constructed to fulfill its role. In that case, we may be able to avoid this failure mode.

This is closely related to interpretability and reducing risks from inner alignment failures:

- If we can limit the amount of black-box compute, and the amount of uninterpretable intermediate state, we're in a better position to know what each of the model components is doing. We view this type of progress as complementary with Chris Olah's work on interpretability and [ELK-style proposals](#) for learning what models know. The better we are at decomposition, the less weight rests on these alternatives.
- Inner alignment failures are most likely in cases where models don't just know a few facts that we don't but can hide extensive knowledge from us, akin to developing new branches of science that we can't follow. With limited compute and limited neural memory, the risk is lower. Advancing process-based systems is helpful on the margin, even if we can't fully eliminate outcome-based optimization.

In the long run, differential capabilities and alignment converge

Today, differential capabilities and alignment look different. Differential capabilities are starting to matter now. Alignment is a much less prominent issue because we don't yet have AI systems that are good at gaming our metrics.

In the crazy future, when automated systems are much more capable and make most decisions in the world, differential capabilities and alignment are two sides of the same coin:

1. We either can't use AI for most tasks we care about if all we know is how to design outcome-based architectures (lack of capabilities), or

2. We have highly effective systems optimizing for flawed objectives, which can lead to catastrophic outcomes (misalignment)

People sometimes ask: Is Ought working on differential capabilities (making ML useful for supporting reasoning) or alignment (avoiding risks from advanced AI)? From the perspective of intervening by advancing process-based systems, these two causes are fundamentally tied together.

Two attractors: The race between process- and outcome-based systems

Outcome-based optimization is an attractor

In some sense, you could almost always do better through end-to-end training, at least according to any one metric. You start with a meaningful task decomposition, track a global metric, and then backpropagate to make the system better along that metric. This messes with the meaning of the components and soon, they can't be interpreted in isolation anymore.

We expect that, at some point, there will be strong pressure to optimize the components of most digital systems we're using for global metrics. The better we are at building process-based systems, the less pressure there will be.

Process-based optimization could be an attractor, too

The good crazy future is one with an ecosystem of AIs made out of components with roles that are in principle human-understandable, with each component optimized based on how well it accomplishes its local role.

Advanced process-based systems could self-regulate to remain process-based, which makes them a local attractor:

1. Whenever an action is chosen within the process-based system, it comes from an action suggester along with reasoning for why it's good for the system to implement this action
2. This suggester could propose to make local changes, like changing some weights, just because empirically they'll improve the quality of overall results along some metric, even if it makes the system less modular and interpretable
3. This proposal and the reasoning for it would then get evaluated by another part of the system that looks for errors and catches and fixes them before they matter
4. This evaluator would evaluate the costs and benefits of implementing the proposal and reject it because it would not maintain the invariant that each component has a clear role that makes sense independent of the global objective

This story makes the basin of attraction around process-based systems look a lot more narrow than the basin around outcomes: It only applies to individual systems, and it

assumes that there is a fairly bright line between components that have a clear role and those that don't.

The state of the race

Today, process-based systems are ahead: Most systems in the world don't use much machine learning, and to the extent that they use it, it's for small, independently meaningful, fairly interpretable steps like predictive search, ranking, or recommendation as part of much larger systems.

However, the history of machine learning is [the bitter lesson](#) of outcomes winning. Vision and NLP started with more structured systems, which were replaced with end-to-end systems. In these areas, the structured systems are much worse, and we don't know how to make them competitive on standard benchmarks. Deepmind and OpenAI have better infrastructure for running RL on outcome-based metrics than for collecting process-based feedback. They tend towards a "research aesthetic" that favors outcomes-based approaches even in cases where they work worse.

Overall, it's up in the air which tasks will be solved in which way. Some parts of the AI community are leaning toward process, others toward outcomes. If we see impressive results from process-based feedback, institutional knowledge and research tastes may shift toward process-based systems. Future norms and laws, perhaps similar to existing algorithmic transparency laws, might strengthen this position.

We don't need process-based systems to be a perfect attractor. If most systems are largely process-based around the time of transformative AI, with small amounts of outcome-based optimization, we're likely in good shape.

Conclusion

If we run into trouble with early advanced AI systems, it will likely be clear that supervision of process would be better than supervision of outcomes. At that point, the question is whether we're good enough at process-based systems that they're a realistic option. If so, then for the most important and high-stakes use cases, people will likely switch. This requires that we develop the relevant know-how now.

Beyond AI, we view understanding how to build systems and institutions that make correct decisions even when outcomes aren't available as part of a broader agenda of advancing reason and wisdom in the world. Making mistakes about the long-term consequences of our short-term decisions is one way we fall short of our potential. Making wise decisions in cases where we can't easily learn from our failures is likely key to living up to it.

Acknowledgments

Thanks to Paul Christiano and Jon Uesato for relevant discussions, and Jon Uesato, Owain Evans, Ben Rachbach, and Luke Stebbing for feedback on a draft.

1. ^

What "crazy" means:

- AI systems are doing most economically valuable tasks in the world. They're developing, producing, and shipping new products. They're writing code, running datacenters, and developing new technologies. They're influencing policy to some extent.
- An increasingly large part of the world economy is AI development, more than shows up explicitly because all fields depend on AI now. The AI industry is worth many trillions of dollars.
- As more of the world economy depends on AI, the value of further improvements to AI increases. It is hard to scale up human researchers and programmers working on AI. Automation of AI research is one of the most important application areas of AI—rolling out AI in new domains, making existing applications better, improving hardware, software, and data centers.
- Much of this activity happens without humans in the loop. It's a complex economy of AI systems.
- This transition to an AI-run economy could be centralized in one or a few firms, or involve many firms, each specializing in different roles. It could take two decades, or five, and the path there could be more or less continuous. Either way, we think it's likely that the world within our lifetime will look very different from today's world in ways that will be obvious to everyone.

Takeoff speeds have a huge effect on what it means to work on AI x-risk

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

The slow takeoff hypothesis predicts that AGI emerges in a world where powerful but non-AGI AI is already a really big deal. Whether AI is a big deal right before the emergence of AGI determines many super basic things about what we should think our current job is. I hadn't fully appreciated the size of this effect until a few days ago.

In particular, in a fast takeoff world, AI takeover risk never looks much more obvious than it does now, and so x-risk-motivated people should be assumed to cause the majority of the research on alignment that happens. In contrast, in a slow takeoff world, many aspects of the AI alignment problems will already have showed up as alignment problems in non-AGI, non-x-risk-causing systems; in that world, there will be lots of industrial work on various aspects of the alignment problem, and so EAs now should think of themselves as trying to look ahead and figure out which margins of the alignment problem aren't going to be taken care of by default, and try to figure out how to help out there.

In the fast takeoff world, we're much more like a normal research field—we want some technical problem to eventually get solved, so we try to solve it. But in the slow takeoff world, we're basically in a weird collaboration across time with the more numerous, non-longtermist AI researchers who will be in charge of aligning their powerful AI systems but who we fear won't be cautious enough in some ways or won't plan ahead in some other ways. Doing technical research in the fast takeoff world basically just requires answering technical questions, while in the slow takeoff world your choices about research projects are closely related to your sociological predictions about what things will be obvious to whom when.

I think that these two perspectives are extremely different, and I think I've historically sometimes had trouble communicating with people who held the slow takeoff perspective because I didn't realize we disagreed on basic questions about the conceptualization of the question. (These miscommunications persisted even after I was mostly persuaded of slow takeoffs, because I hadn't realized the extent to which I was implicitly assuming fast takeoffs in my picture of how AGI was going to happen.)

As an example of this, I think I was quite confused about what genre of work various prosaic alignment researchers think they're doing when they talk about alignment schemes. To quote [a recent AF shortform post of mine](#):

Something I think I've been historically wrong about:

A bunch of the prosaic alignment ideas (eg adversarial training, IDA, debate) now feel to me like things that people will obviously do the simple versions of by default. Like, when we're training systems to answer questions, of course we'll use our current versions of systems to help us evaluate, why would we not do that? We'll be used to using these systems to answer questions that we have, and so it will be totally obvious that we should use them to help us evaluate our new system.

Similarly with debate--adversarial setups are pretty obvious and easy.

In this frame, the contributions from Paul and Geoffrey feel more like “they tried to systematically think through the natural limits of the things people will do” than “they thought of an approach that non-alignment-obsessed people would never have thought of or used”.

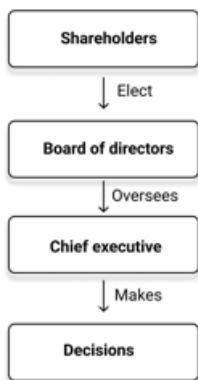
It’s still not obvious whether people will actually use these techniques to their limits, but it would be surprising if they weren’t used at all.

I think the slightly exaggerated slogan for this update of mine is “IDA is futurism, not a proposal”.

My current favorite example of the thinking-on-the-margin version of alignment research strategy is in [this comment by Paul Christiano](#).

Ideal governance (for companies, countries and more)

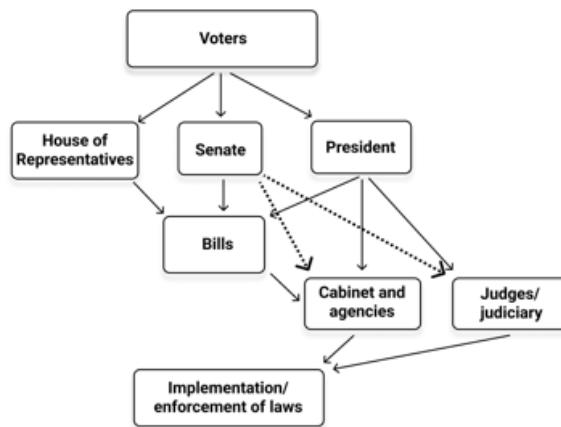
Corporation:



Nonprofit:



US Federal Govt:



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

I'm interested in the topic of **ideal governance**: what kind of governance system should you set up, if you're starting from scratch and can do it however you want?

Here "you" could be a company, a nonprofit, an informal association, or a country. And "governance system" means a Constitution, charter, and/or bylaws answering questions like: "Who has the authority to make decisions (Congress, board of directors, etc.), and how are they selected, and what rules do they have to follow, and what's the process for changing those rules?"

I think this is a very different topic from something like "How does the US's Presidential system compare to the Parliamentary systems common in Europe?" The idea is not to look at today's most common systems and compare them, but rather to generate options for setting up systems radically different from what's common today.

I don't currently know of much literature on this topic (aside from the literature on [social choice theory](#) and especially [voting methods](#), which covers only part of the topic). This post **describes the general topic and why I care, partly in the hopes that people can point me to any literature I've missed**. Whether or not I end up finding any, I'm likely to write more on this topic in the future.

Outline of the rest of the piece:

- I'll outline some common governance structures for countries and major organizations today, and highlight **how much room there is to try different things** that don't seem to be in wide use today. [More](#)
- I'll discuss **why I care about this question**. I have a few very different reasons:

- A *short-term, tangible need*: over the last several years, I've spoken with several (more than 3) organizations that feel no traditional corporate governance structure is satisfactory, because the stakes of their business are too great and society-wide for shareholder control to make sense, yet they are too early-stage and niche (and in need of nimbleness) to be structured like a traditional government. An example would be an artificial intelligence company that could end up with a normal commercial product, or could end up bringing about the [most important century of all time for humanity](#). I wish I could point them to someone who was like: "I've read all of, and written much of, the literature on what your options are. I can walk you through the pros and cons and help you pick a governance system that balances them for your needs."
- A *small probability of a big future win*. The world today has lots of governments, but they seem to mostly follow a very small number of basic governance templates. At some point, there will be new states with new Constitutions - maybe via space settlements, maybe via collapse of existing states, etc. - but I expect these moments to be few and far between. A significant literature and set of experts on "ideal governance" could lead to a radically different kind of state government, potentially with radically different policies that the rest of the world could learn from.
- A *weird, out-of-left-field application*. Some of my interest in this topic actually comes via my interest in [moral uncertainty](#): the question of what it's ethical to do when one is *struggling between more than one theory of ethics*, with radically different implications. This is hard to explain, but I try [below](#).
- I'll describe a bit more what I think **literature on this question could look like** (and what already exists that I know of), partly to guide readers who might be able to help me find more.

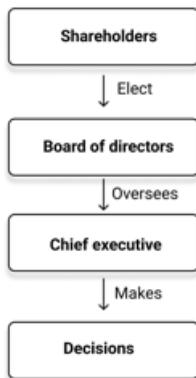
Common governance structures today

All of these are simplified; I'm trying to illustrate the basic idea of what questions "ideal governance" is asking.

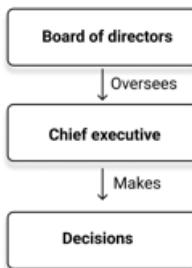
- A standard (e.g., public) corporation works like this: it has shareholders, assigned one vote per share (not per person), who elect a board of directors that governs by majority. The board generally appoints a CEO that it entrusts with day-to-day decisions. There is a "constitution" of sorts (the Articles of Incorporation and bylaws) and a lot more wrinkles in terms of how directors are selected, but that's the basic idea.
- A standard nonprofit is like a corporation, but entirely lacking the shareholder layer - it's governed directly by the board of directors. (I find something weird about a structure this simple - a simple board majority can do literally anything, even though the board of directors is often a somewhat random assortment of donors, advisors, etc.)
- The US federal government is a lot more complex. It splits authority between the House of Representatives, the Senate, the Presidency and the Supreme Court, all of which have specific appointment procedures, term limits, etc. and are meta-governed by a Constitution that requires special measures to change. There are lots of specific choices that were made in designing things this way, and lots of things that could've been set up differently in the 18th century that would probably still matter today.

- Other democracies tend to have governments that differ in a lot of ways ([e.g.](#)), while being based on broadly similar principles: voters elect representatives to more than one branch of government, which then divide up (and often can veto each other on) laws, expenditures, etc.
- When I was 13, the lunch table I sat at established a Constitution with some really strange properties that I can't remember. I think there was a near-dictatorial authority who rotated daily, with others able to veto their decisions by assembling supermajorities or maybe singing silly songs or something.

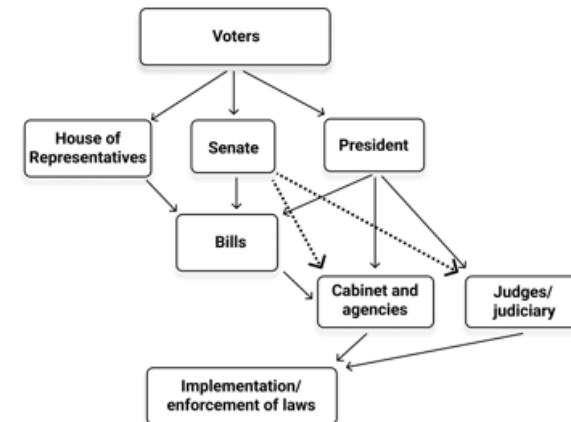
Corporation:



Nonprofit:



US Federal Govt:



In addition to the design choices shown in the diagrams, there are a lot of others:

- Who votes, how often, and what voting system is used?
- How many representatives are there in each representative body? How are they divided up (one representative per geographic area, or [party-list proportional representation](#), or something else)?
- What term limits exist for the different entities?
- Do particular kinds of decisions require supermajorities?
- Which restrictions are enshrined in a hard-to-change Constitution (and how hard is it to change), vs. being left to the people in power at the moment?

One way of thinking about the "ideal governance" question is: **what kinds of designs could exist that aren't common today? And how should a new organization/country/etc. think about what design is going to be best for its purposes, beyond "doing what's usually done"?**

For any new institution, it seems like the stakes are potentially high - in some important sense, picking a governance system is a "one-time thing" (any further changes have to be made using the rules of the existing system¹).

Perhaps because of this, there doesn't seem to be much use of innovative governance designs in high-stakes settings. For example, here are a number of ideas I've seen floating around that seem cool and interesting, and ought to be considered if someone could set up a governance system however they wanted:

- **Sortition**, or choosing people randomly to have certain powers and responsibilities. An extreme version could be: "Instead of everyone voting for President, randomly select 1000 Americans; give them several months to

consider their choice, perhaps paid so they can do so full-time; then have them vote."

- The idea is to pick a subset of people who are both (a) representative of the larger population (hence the randomness); (b) will have a stronger case for putting serious time and thought into their decisions (hence the small number).
- It's solving a similar problem that "representative democracy" (voters elect representatives) is trying to solve, but in a different way.
- **Proportional decision-making.** Currently, if Congress is deciding how to spend \$1 trillion, a coalition controlling 51% of the votes can control all \$1 trillion, whereas a coalition controlling 49% of the votes controls \$0. Proportional decision-making could be implemented as "Each representative controls an equal proportion of the spending," so a coalition with 20% of the votes controls 20% of the budget. It's less clear how to apply this idea to other sorts of bills (e.g., illegalizing an activity rather than spending money), but there are plenty of possibilities.²
- **Quadratic voting**, in which people vote on multiple things at once, and can cast more votes for things they care about more (with a "quadratic pricing rule" intended to make the number of votes an "honest signal" of how much someone cares).
- **Reset/Jubilee:** maybe it would be good for some organizations to periodically redo their governance mostly from scratch, subject only to the most basic principles. Constitutions could contain a provision like "Every N years, there shall be a new Constitution selected. The 10 candidate Constitutions with the most signatures shall be presented on a ballot; the Constitution receiving the most votes is the new Constitution, except that it may not contradict or nullify this provision. This provision can be prevented from occurring by [supermajority provision], and removed entirely by [stronger supermajority]."
- More examples in a footnote.³

If we were starting a country or company from scratch, which of the above ideas should we integrate with more traditional structures, and how, and what else should we have in our toolbox? That's the question of ideal governance.

Why do I care?

I have one "short-term, tangible need" reason; one "small probability of a big future win" reason; and one "weird, out-of-left-field" reason.

A short-term, tangible need: companies developing AI, or otherwise aiming to be working with huge stakes. Say you're starting a new company for developing AI systems, and you believe that you could end up building [AI with the potential to change the world forever](#).

- The standard governance setup for a corporation would hand power over all the decisions you're going to make to your shareholders, and by default most of your shares are going to end up held by people and firms that invested money in your company. Hopefully it's clear why this doesn't seem like the ideal setup for a company whose decisions could be world-changing. A number of AI companies have acknowledged the basic point that "Our ultimate mission should NOT just be: make money for shareholders," and that seems like a good thing.
- One alternative would be to set up like a nonprofit instead, with all power vested in a board of directors (no shareholder control). Some issues are that (a) this cuts

shareholders out of the loop completely, which could make it pretty hard to raise money; (b) according to me at least, this is just a weird system of governance, for reasons that are not super easy to articulate concisely but I'll take a shot in a footnote⁴ (and possibly write more in the future).

- Another alternative is a setup that is somewhat common among tech companies: 1-2 founders hold enough shares to keep control forever, so you end up with essentially a dictatorship. This also ... leaves something to be desired.
- Or maybe a company like this should just set up more like a government from the get-go, offering everyone in the world a vote via some complex system of representation, checks and balances. But this seems poorly suited to at least the relatively early days of a company, when it's small and its work is not widely known or understood. But then, how does the company handle the transition from the latter to the former? And should the former be done exactly in the standard way, or is there room for innovation there?

Over the last several years, I've spoken with heads of several (more than 3) organizations that struggle between options like the above, and have at least strongly considered unusual governance setups. I wish I could point them to someone who was like: "I've read all of, and written much of, the literature on what your options are. I can walk you through the pros and cons and help you pick a governance system that balances them for your needs."

But right now, I can't, and I've seen a fair amount of this instead: "Let's just throw together the best system we can, based mostly on what's already common but with a few wrinkles, and hope that we figure this all out later." I think this is the right solution given how things stand, but I think it really does get continually harder to redesign one's governance as time goes on and more [stakeholders](#) enter the picture, so it makes me nervous.

Similar issues could apply to mega-corporations (e.g., [FAANG](#)) that are arguably more powerful than what the standard shareholder-centric company setup was designed for. Are there governance systems they could adopt that would make them more broadly accountable, without copying over all the pros and cons of full-blown representative democracy as implemented by countries like the US?

A small probability of a big future win: future new states. The world today has lots of governments, but they seem to mostly follow a very small number of basic governance templates (e.g., I believe you see almost none of [the things I listed above](#)), and probably relatedly, there seems to be remarkably little variety and experimentation with policy. Policies that many believe could be huge wins - such as [dramatically expanded immigration](#), [land value taxation](#), "consumer reports"-style medical approvals,⁵ [drug decriminalization](#), and [charter cities](#) - don't seem to have gotten much of a trial anywhere in the world.

At some point, there will be new states with new Constitutions - maybe via space settlements, maybe via collapse of existing states, etc. - but I expect these moments to be few and far between.

By default I expect future Constitutions to resemble present ones an awful lot. But maybe, at some future date, there will be a large "ideal governance" literature and some points of expert consensus on innovative governance designs that somebody really ought to try. That could lead to a radically different kind of state government, potentially with radically different policies that the rest of the world could learn from.

An out-of-left-field application for "ideal governance." This is going to veer off the rails, so remember to skip to the next section if I lose you.

Some of my interest in this topic actually comes via my interest in [moral uncertainty](#): the question of what it's ethical to do when one is *struggling between more than one theory of ethics*, with radically different implications.

For example, there are arguments that our [ethical decisions should be dominated by concern for ensuring that as many people as possible will someday get to exist](#). I really go back and forth on how much I buy these arguments, but I'm definitely somewhere between 10% convinced and 50% convinced. So ... say I'm "20% convinced" of some view that says preventing human extinction⁶ is the [overwhelmingly](#) most important consideration for at least some dimensions of ethics (like where to donate), and "80% convinced" of some more common-sense view that says I should focus on some cause unrelated to human extinction.⁷ How do I put those two together and decide what this means for actual choices I'm making?

The closest thing I've seen to a reasonable-seeming answer is the idea of a [moral parliament](#): I should act as though I'm run by a Parliament with 80 members who believe in "common-sense" ethics, and 20 members who believe in the "preventing extinction is overwhelmingly important" idea. But with default Parliament rules, this would just mean the 80 members can run the whole show, without any compromise with the 20.

And so, a [paper on the "moral parliament" idea](#) tries to make it work by ... introducing a completely new governance mechanism that I can't find any other sign of someone else ever talking about, "proportional chances voting" (spelled out in a footnote).⁸ I think this mechanism has its own issues,⁹ but it's an attempt to ensure something like "A coalition controlling 20% of the votes has 20% of the effective power, and has to be compromised with, instead of being subject to the tyranny of the majority."

My own view (which I expect to write more about in the future) is that governance is roughly the right metaphor for "moral uncertainty": I am torn by [multiple different sides of myself](#), with different takes on what it means to be a good person, and the problem of getting these different sides of myself to reach a decision together is like the problem of getting different citizens (or shareholders) to reach a decision together. The more we can say about what ideal governance looks like, the more we can say about how this ought to work - and the better I expect this "moral parliament"-type idea to end up looking, compared to alternatives.¹⁰

The literature I'm looking for

Ideal governance seems like the sort of topic for which there should be a "field" of "experts," studying it. What would such study look like? Three major categories come to mind:

Brainstorming ideas such as [those I listed above](#) - innovative potential ways of solving classic challenges of governance, such as reconciling "We want to represent all the voters" with "We want decisions to be grounded in expertise and high engagement, and voters are often non-expert and not engaged."

I've come across various assorted ideas in this category, including [quadratic voting](#), [futarchy](#), and [proportional chances voting](#), without seeing much sign that these sit

within a broader field that I can skim through to find all the ideas that are out there.

Economics-style theory in which one asks questions like: "If we make particular assumptions about who's voting, what information they have and lack, how much they suffer from [bounded rationality](#), and how we define 'serving their interests' (see below), what kind of governance structure gets the best outcome?"

[Social choice theory](#), including on [voting methods](#), tackles the "how we define 'serving their interests'" part of this. But I'm not aware of people using similar approaches to ask questions like "Under what conditions would we want 1 chamber of Congress vs. 2, or 10? 100 Senators vs. 500, or 15? A constitution that can be modified by simple majority, vs. 2/3 majority vs. consensus? Term limits? Etc. etc. etc."

Empirical research (probably qualitative): Are there systematic reviews of unusual governance structures tried out by companies, and what the results have been? Of smaller-scale experiments at co-ops, group houses and lunch tables?

To be clear, I think the most useful version of this sort of research would probably be very qualitative - collecting reports of what problems did and didn't come up - rather than asking questions like "How does a particular board structure element affect company profits?"

One of the things I expect to be tricky about this sort of research is that I think a lot of governance comes down to things like "What sorts of people are in charge?" and "What are the culture, expectations, norms and habits?" A setup that is "officially" supposed to work one way could evolve into something quite different via informal practices and "soft power." However, I think the formal setup (including things like "what the constitution says about the principles each governance body is supposed to be upholding") can have big effects on how the "soft power" works.

If you know where to find research or experts along the lines of the above, please share them [in the comments](#) or [using this form](#) if you don't want them to be public.

I'll likely write about what I come across, and if I don't find anything new, I'll likely ramble some more about ideal governance. So either way, there will be more on this topic!









[Comment/discuss](#)

Footnotes

1. Barring violent revolution in the case of countries. [←](#)
2. An example would be the "proportional chances voting" idea described [here](#). [←](#)
3.
 - **Proxying/liquid democracy**, or allowing voters to transfer their votes to other voters. (This is common for corporations, but not for governments.) This could be an alternative or complement to electing representatives, solving a similar problem (we want lightly-engaged voters to be

represented, but we also want decisions ultimately made using heavy engagement and expertise). At first glance it may seem to pose a risk that people will be able to "buy votes," but I don't actually think this is necessarily an issue (proxying could be done anonymously and on set schedules, like other votes).

- **Soft term limits:** the more terms someone has served, the greater a supermajority they need to be re-elected. This could be used to strike a balance between the advantages of term limits (avoiding "effectively unaccountable" incumbents) and no-term-limits (allowing great representatives to keep serving).
 - **Formal technocracy/meritocracy:** Using hard structures (rather than soft norms) to assign authority to people with particular expertise and qualifications. An extreme example would be [futarchy](#), in which prediction markets directly control decisions. A simpler example would be structurally rewarding representatives (via more votes or other powers) based on assessments of their track records (of predictions or decisions), or factual understanding of a subject. This seems like a tough road to go down by default, as any mechanism for evaluating "track records" and "understanding" can itself be politicized, but there's a wide space of possible designs. ↩
4. Most systems of government have a sort of funnel from "least engaged in day to day decisions, but most ultimately legitimate representatives of whom the institution is supposed to serve" (shareholders, voters) to "most engaged in day to day decisions, but ultimately accountable to someone else" (chief executive). A nonprofit structure is a very short funnel, and the board of directors tends to be a somewhat random assortment of funders, advisors, people who the founders just thought were cool, etc. I think they often end up not very accountable (to anyone) or engaged in what's going on, such that they have a hard time acting when they ought to, and the actions they do take are often kind of random.
- I'm not saying there is a clearly better structure available for this purpose - I think the weirdness comes from the fact that it's so unclear who should go in the box normally reserved for "Shareholders" or "Voters." It's probably the best common structure for its purpose, but I think there's a lot of room for improvement, and the stakes seem high for certain organizations. ↩
5. Context in [this Marginal Revolution post](#), which links to [this 2005 piece on a "consumer reports" model for the FDA](#). ↩
6. Or "existential catastrophe" - something that drastically curtails humanity's future, even if it doesn't drive us extinct. ↩
7. This isn't actually where I'm at, because I think the leading existential risks are a big enough deal that I would want to focus on them even if I completely ignored the philosophical argument that the future is overwhelmingly important. ↩
8. Let's say that 70% of the Parliament members vote for bill X, and 30% vote against. "Proportional chance voting" literally uses a weighted lottery to pass bill X with 70% probability, and reject it with 30% probability (you can think of this like rolling a 10-sided die, and passing the bill if it's 7 or under).

A key part of this is that the members are supposed to negotiate before voting and holding the lottery. For example, maybe 10 of the 30 members who are against bill X offer to switch to supporting it if some change is made. The nice

property here is that rather than having a "tyranny of the majority" where the minority has no bargaining power, we have a situation where the 70-member coalition would still love to make a deal with folks in the minority, to further increase the probability that they get their way.

Quote from the paper that I am interpreting: "Under proportional chances voting, each delegate receives a single vote on each motion. Before they vote, there is a period during which delegates may negotiate: this could include trading votes on one motion for votes on another, introducing novel options for consideration within a given motion, or forming deals with others to vote for a compromise option that both consider to be acceptable. The delegates then cast their ballots for one particular option in each motion, just as they might in a plurality voting system. But rather than determining the winning option to be the one with the most votes, each option is given a chance of winning proportional to its share of the votes." [←](#)

9. What stops someone who lost the randomized draw from just asking to hold the same vote again? Or asking to hold a highly similar/related vote that would get back a lot of what they lost? How does that affect the negotiated equilibrium? [←](#)
10. Such as "[maximize expected choice-worthiness](#)," which I am not a fan of for reasons I'll get to in the future. [←](#)

Convincing All Capability Researchers

This post is heavily based off this [excellent comment](#) (the author's name is not relevant)

Give the world's thousand most respected AI researchers \$1M each to spend 3 months working on AI alignment, with an extra \$100M if by the end they can propose a solution alignment researchers can't shoot down. I promise you that other than like 20 industry researchers who are paid silly amounts, every one of them would take the million. They probably won't make any progress, but from then on when others ask them whether they think alignment is a real unsolved problem, they will be way more likely to say yes. That only costs you a billion dollars! I literally think I could get someone reading this the money to do this (at least at an initially moderate scale) - all it needs is a competent person to step up.

Usually people object to paying people large sums of money to work on alignment because they don't expect them to produce any good work (mostly because it's very hard to specify alignment, see below). This is a feature, not a bug.

Being able to say "the 1000 smartest people working in AI couldn't make headway in Alignment in 3 months even when they were paid \$1 million and a solution would be awarded \$100 million" is very good for persuading existing researchers that this is a very hard problem.

Would this really go a long way to convincing those at major AI labs that alignment is hard? We could actually ask people working at these places if there were no progress made in those 3 months if it would change their minds.

Another problem is looooong timelines. They can agree it's hard, but their timelines may be too far to matter to work on it now. I could think of a couple counter-arguments that may work (but it would better to actually test them with real people)

- 1. To predict AGI (or just AI that has the capability of destroying the world), you could argue "It needs these 10 capabilities, and I predict it will get each capabilities at these different time periods". If someone actually did that and was historically correct, even within 5-10 years, I'd be impressed and trust their opinion. Have you successfully predicted AI capabilities of the past 10 years?
 - possibly mention most recent advances (Transformers, some RL, Neural Nets in 2010's)
- 2. If there's a 10% chance that it happens during you or your kids lifetime, then that's extremely important. For example, if I thought there was a 10% chance that I died taking the train home tonight, I wouldn't take the train. Even with a 0.1%, I wouldn't.

Though, generally having a set of people working on apologetics for both US & Chinese AI Researchers is potentially extremely high-impact. If you're interested and might be able to donate, DM me for a call.

Next Steps to seriously consider this proposal

Specifying "Solving Alignment"

Part of the problem of alignment is we don't know the correct framework to specify it. The quoted text suggests the criteria for a solution as "a solution alignment researchers can't shoot down.", which side-steps this issue; however, specifying the problem in as fine-grained detail would be extremely useful for communicating to these researchers.

One failure mode would be them taking the money, not get work done, and then argue the problem wasn't specified enough to make any meaningful progress which limits how persuasive this stunt could be. Documents like ELK are more useful specifications that captures the problem to various degrees, and I wish we had more problems like that.

Listing 1000 AI researcher intellectuals

The initial idea is to hire 1000 best AI researchers to work on the problem, not because we expect them to solve it, but by all of them failing

here are a few different proxy's we can use, such as citations and top researchers at AI Labs. So far I've got

- 1. [Authors with the most citations in ML from google scholar](#)
- 2. DeepMind, Facebook, Microsoft, Tesla, OpenAI, major university groups(?) , ?
- 3. ~~Myself and all my friends~~
- [3.Every author on this Chinese paper](#)

Convincing the CCP and backed researchers is a blank spot in my map, and if anyone knows anything, please comment or message me for a video call.

Actually Hiring People/Creating a Company to Do This

We would need competent people to work on the specification of the problem, outreach and selecting who to pay, keeping up with researchers (if you're paying them \$1 million dollars, you also can have 1-on-1 calls which would be useful to make the most of), and reviewing actual work produced(which could actually be done by the community/ independent researchers/orgs).

Timelines Argument

It was argued that these plans are only relevant in 15+ timelines, but huge social changes/cultural norms have happened within 1 year time periods. I'm not giving examples, but they went from, say 50 to 100 to 70 within a few months, which may be significantly different than Alignment.

This Plan May Backfire and Increase Capabilities

A way this can backfire is increasing race conditions, such that everyone wants to create the AGI first. Or at least, more people than are already doing so right now.

I think this is a relevant possibility, and this should be taken into account with whoever reaches out to talk to these top researchers when offering to pay them the large sums of money.

Google's new 540 billion parameter language model

This is a linkpost for <https://storage.googleapis.com/pathways-language-model/PaLM-paper.pdf>

Google just announced a very large language model that achieves SOTA across a very large set of tasks, mere days after DeepMind [announced](#) *Chinchilla*, and [their discovery](#) that data-scaling might be more valuable than we thought.

Here's [the blog post](#), and here's [the paper](#). I'll repeat the abstract here, with a highlight in bold,

Large language models have been shown to achieve remarkable performance across a variety of natural language tasks using few-shot learning, which drastically reduces the number of task-specific training examples needed to adapt the model to a particular application. To further our understanding of the impact of scale on few-shot learning, we trained a 540-billion parameter, densely activated, Transformer language model, which we call Pathways Language Model (PaLM).

We trained PaLM on 6144 TPU v4 chips using Pathways, a new ML system which enables highly efficient training across multiple TPU Pods. We demonstrate continued benefits of scaling by achieving state-of-the-art few-shot learning results on hundreds of language understanding and generation benchmarks. On a number of these tasks, PaLM 540B achieves breakthrough performance, outperforming the finetuned state-of-the-art on a suite of multi-step reasoning tasks, and outperforming average human performance on the recently released BIG-bench benchmark. **A significant number of BIG-bench tasks showed discontinuous improvements from model scale, meaning that performance steeply increased as we scaled to our largest model.** PaLM also has strong capabilities in multilingual tasks and source code generation, which we demonstrate on a wide array of benchmarks. We additionally provide a comprehensive analysis on bias and toxicity, and study the extent of training data memorization with respect to model scale. Finally, we discuss the ethical considerations related to large language models and discuss potential mitigation strategies.

“Pivotal Act” Intentions: Negative Consequences and Fallacious Arguments

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

tl;dr: I know a bunch of EA/rationality-adjacent people who argue — sometimes jokingly and sometimes seriously — that the only way or best way to reduce existential risk is to enable an “aligned” AGI development team to forcibly (even if nonviolently) shut down all other AGI projects, using safe AGI. I find that the arguments for this conclusion are flawed, and that the conclusion itself causes harm to institutions who espouse it. Fortunately (according to me), successful AI labs do not seem to espouse this “pivotal act” philosophy.

[This post is also available on the [EA Forum](#).]

How to read this post

Please read Part 1 first if you’re very impact-oriented and want to think about the consequences of various institutional policies more than the arguments that lead to the policies; then Parts 2 and 3.

Please read Part 2 first if you mostly want to evaluate policies based on the arguments behind them; then Parts 1 and 3.

I think all parts of this post are worth reading, but depending on who you are, I think you could be quite put off if you read the wrong part first and start feeling like I’m basing my argument too much on kinds-of-thinking that policy arguments should not be based on.

Part 1: Negative Consequences of Pivotal Act Intentions

Imagine it's 2022 (it is!), and your plan for reducing existential risk is to build or maintain an institution that aims to find a way for you — or someone else you'll later identify and ally with — to use AGI to forcibly shut down all other AGI projects in the world. By "forcibly" I mean methods that violate or threaten to violate private property or public communication norms, such as by using an AGI to engage in...

- **cyber sabotage**: hacking into competitors' computer systems and destroy their data;
- **physical sabotage**: deploying tiny robotic systems that locate and destroy AI-critical hardware without (directly) harming any humans;
- **social sabotage**: auto-generating mass media campaigns to shut down competitor companies by legal means, or
- **threats**: demonstrating powerful cyber or physical or social threats, and bargaining with competitors to shut down "or else".

Hiring people for your pivotal act project is going to be tricky. You're going to need people who are willing to take on, or at least tolerate, a highly adversarial stance toward the rest of the world. I think this is very likely to have a number of bad consequences for your plan to do good, including the following:

1. (**bad external relations**) People on your team will have a low trust and/or adversarial stance towards neighboring institutions and collaborators, and will have a hard time forming good-faith collaboration. This will alienate other institutions and make them not want to work with you or be supportive of you.
2. (**bad internal relations**) As your team grows, not everyone will know each other very well. The "us against the world" attitude will be hard to maintain, because there will be an ever weakening sense of "us", especially as people quit and move to other institutions and conversely. Sometimes, new hires will express opinions that differ from the dominant institutional narrative, which might pattern-match as "outsidery" or "norm-y" or "too caught up in external politics", triggering feelings of internal distrust within the team that some people might defect on the plan to forcibly shut down other projects. This will cause your team to get along poorly internally, and make it hard to manage people.
3. (**risky behavior**) In the fortunate-according-to-you event that your team manages to someday wield a powerful technology, there will be a sense of pressure to use it to "finally make a difference" or other argument that boils down to acting quickly before competitors would have a chance to shut you down or at least defend themselves. This will make it hard to stop your team from doing rash things that would actually increase existential risk.

Overall, building an AGI development team with the intention to carry out a "pivotal act" of the form "forcibly shut down all other A(G)I projects" is probably going to be a rough time, I predict.

Does this mean no institution in the world can have the job of preparing to shut down runaway technologies? No; see "Part 3: it matters who does things".

Part 2: Fallacies in Justifying Pivotal Acts

For pivotal acts of the form “shut down all (other) AGI projects”, there’s an argument that I’ve heard repeatedly from dozens of people, which I claim has easy-to-see flaws if you slow down and visualize the world that the argument is describing.

This is *not* an argument that successful AI research groups (e.g., OpenAI, DeepMind, Anthropic) seem to espouse. Nonetheless, I hear the argument frequently enough to want to break it down and refute it.

Here is the argument:

1. AGI is a dangerous technology that could cause human extinction if not super-carefully aligned with human values.

(My take: I agree with this point.)

2. If the first group to develop AGI manages to develop safe AGI, but the group allows other AGI projects elsewhere in the world to keep running, then one of those other projects will likely eventually develop unsafe AGI that causes human extinction.

(My take: I also agree with this point, except that I would bid to replace “the group allows” with “the world allows”, for reasons that will hopefully become clear in Part 3: It Matters Who Does Things.)

3. Therefore, the first group to develop AGI, assuming they manage to align it well enough with their own values that they believe they can safely issue instructions to it, should use their AGI to build offensive capabilities for targeting and destroying the hardware resources of other AGI development groups, e.g., nanotechnology targeting GPUs, drones carrying tiny EMP charges, or similar.

(My take: I do not agree with this conclusion, I do not agree that (1) and (2) imply it, and I feel relieved that every successful AI research group I talk to is also not convinced by this argument.)

The short reason why (1) and (2) do not imply (3) is that when you have AGI, you don’t have to use the AGI directly to shut down other projects.

In fact, before you get to AGI, your company will probably develop other surprising capabilities, and you can demonstrate those capabilities to neutral-but-influential outsiders who previously did not believe those capabilities were possible or concerning. In other words, outsiders can start to help you implement helpful regulatory ideas, rather than you planning to do it all on your own by force at the last minute using a super-powerful AI system.

To be clear, I’m not arguing for leaving regulatory efforts entirely in the hands of governments with no help or advice or infrastructural contributions from the tech sector. I’m just saying that there are many viable options for regulating AI technology

without requiring one company or lab to do all the work or even make all the judgment calls.

Q: Surely they must be joking or this must be straw-manning... right?

A: I realize that lots of EA/R folks are thinking about AI regulation in a very nuanced and politically measured way, which is great. And, I don't think the argument (1-3) above represents a majority opinion among the EA/R communities. Still, some people mean it, and more people joke about it in an ambiguous way that doesn't obviously distinguish them from meaning it:

- (**ambiguous joking**) I've numerous times met people at EA/R events who were saying extreme-sounding things like "[AI lab] should just melt all the chip fabs as soon as they get AGI", who when pressed about the extremeness of this idea will respond with something like "Of course I don't *actually mean* I want [some AI lab] to melt all the chip fabs". Presumably, some of those people were actually just using hyperbole to make conversations more interesting or exciting or funny.

Part of my motivation in writing this post is to help cut down on the amount of ambiguous joking about such proposals. As the development of more and more advanced AI technologies is becoming a reality, ambiguous joking about such plans has the potential to really freak people out if they don't realize you're exaggerating.

- (**meaning it**) I have met at least a dozen people who were *not* joking when advocating for invasive pivotal acts along the lines of the argument (1-3) above. That is to say, when pressed after saying something like (1-3), their response wasn't "Geez, I was joking", but rather, "Of course AGI labs should shut down other AGI labs; it's the only morally right thing for them to do, given that AGI labs are bad. And of course they should do it by force, because otherwise it won't get done."

In most cases, folks with these viewpoints seemed not to have thought about the cultural consequences of AGI research labs harboring such intentions over a period of years (Part 2), or the fallacy of assuming technologists will have to do everything themselves (Part 1), or the future possibility of making evidence available to support global regulatory efforts from a broader base of consensual actors (see Part 3).

So, part of my motivation in writing this post is as a genuine critique of a genuinely expressed position.

Part 3: It Matters Who Does Things

I think it's important to separate the following two ideas:

- **Idea A (for “Alright”):** Humanity should develop hardware-destroying capabilities — e.g., broadly and rapidly deployable non-nuclear EMPs — to be used in emergencies to shut down potentially-out-of-control AGI situations, such as an AGI that has leaked onto the internet, or an irresponsible nation developing AGI unsafely.
- **Idea B (for “Bad”):** AGI development teams should be the ones planning to build the hardware-destroying capabilities in Idea A.

For what it's worth, I agree with Idea A, but disagree with Idea B:

Why I agree with Idea A

It's indeed much nicer to shut down runaway AI technologies (if they happen) using hardware-specific interventions than attacks with big splash effects like explosives or brainwashing campaigns. I think this is the main reason well-intentioned people end up arriving at this idea, and Idea B, but I think Idea B has some serious problems.

Why I disagree with Idea B

A few reasons! First, there's:

- **Action Consequence 1:** the action of having an AGI carry out or even prescribe such a large intervention on the world — invading others' private property to destroy their hardware — is risky and legitimately scary. Invasive behavior is risky and threatening enough as it is; using AGI to do it introduces a whole range of other uncertainties, not least because the AGI could be deceptive or otherwise misaligned with humanity in ways that we don't understand.

Second, before even reaching the point of taking the action prescribed in Idea B, merely *harboring the intention* of Idea B has bad consequences; echoing similar concerns as Part 1:

- **Intention Consequence 1: Racing.** Harboring Idea B creates an adversarial winner-takes-all relationship with other AGI companies racing to maintain
 - a degree of control over the future, and
 - the ability to implement their own pet theories on how safety/alignment should work, leading to more desperation, more risk-taking, and less safety overall.
- **Intention Consequence 2: Fear.** Via staff turnover and other channels, harboring Idea B signals to other AGI companies that you are willing to violate their property boundaries to achieve your goals, which will cause them to fear for their physical safety (e.g., because your incursion to invade their hardware might go awry and end up harming them personally as well). This kind of fear leads to more desperation, more winner-takes-all mentality, more risk-taking, and less safety.

Summary

In Part 1, I argued that there are negative consequences to AGI companies harboring the intention to forcibly shut down other AGI companies. In Part 2, I analyzed a common argument in favor of that kind of “pivotal act”, and found a pretty simple flaw stemming from fallaciously assuming that the AGI company has to do everything itself (rather than enlisting help from neutral outsiders, using evidence). In Part 3, I elaborated more on the nuance regarding who (if anyone) should be responsible for developing hardware-shutdown technologies to protect humanity from runaway AI disasters, and why in particular AGI companies should not be the ones planning to do this, mostly echoing points from Part 1.

Fortunately, successful AI labs like DeepMind, OpenAI, and Anthropic do not seem to espouse this “pivotal act” philosophy for doing good in the world. One of my hopes in writing this post is to help more EA/R folks understand why I agree with their position.

A broad basin of attraction around human values?

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

Followup to: [Morality is Scary, AI design as opportunity and obligation to address human safety problems](#)

In [Corrigibility](#), Paul Christiano argued that in contrast with ambitious value learning, an act-based corrigible agent is safer because there is a broad basin of attraction around corrigibility:

In general, an agent will prefer to build other agents that share its preferences. So if an agent inherits a distorted version of the overseer's preferences, we might expect that distortion to persist (or to drift further if subsequent agents also fail to pass on their values correctly).

But a corrigible agent prefers to build other agents that share the overseer's preferences—even if the agent doesn't yet share the overseer's preferences perfectly. After all, even if you only approximately know the overseer's preferences, you know that the overseer would prefer the approximation get better rather than worse.

Thus an entire neighborhood of possible preferences lead the agent towards the same basin of attraction. We just have to get “close enough” that we are corrigible, we don't need to build an agent which exactly shares humanity's values, philosophical views, or so on.

But it occurs to me that the overseer, or the system composing of overseer and corrigible AI, itself constitutes an agent with a distorted version of the overseer's true or actual preferences (assuming a metaethics in which this makes sense, i.e., where one can be wrong about one's values). Some possible examples of human overseer's distorted preferences, in case it's not clear what I have in mind:

1. Wrong object level preferences, such as overweighting values from a contemporary religion or ideology, and underweighting other plausible or likely moral concerns.
2. Wrong meta level preferences (preferences that directly or indirectly influence one's future preferences), such as lack of interest in finding or listening to arguments against one's current moral beliefs, willingness to use "cancel culture" and other coercive persuasion methods against people with different moral beliefs, awarding social status for moral certainty instead of uncertainty, and the revealed preferences of many powerful people for advisors who reinforce one's existing beliefs instead of critical or neutral advisors.
3. Ignorance / innocent mistakes / insufficiently caution meta level preferences in the face of dangerous new situations. For example, what kinds of experiences (especially exotic experiences enabled by powerful AI) are safe or benign to have, what kinds of self-modifications to make, what kinds of people/AI to surround oneself with, how to deal with messages that are potentially AI-optimized for persuasion.

In order to conclude that a corrigible AI is safe, one seemingly has to argue or assume that there is a broad basin of attraction around the overseer's true/actual values (in addition to around corrigibility) that allows the human-AI system to converge to correct values despite starting with distorted values. But if there actually was a broad basin of attraction around human values, then "we don't need to build an agent which exactly shares humanity's values, philosophical views, or so on" could apply to other alignment approaches besides corrigibility / intent alignment, such as ambitious value learning, thus undermining Paul's argument in "Corrigibility". One immediate upshot seems to be that I, and others who were persuaded by that argument, should perhaps pay a bit more attention to other approaches.

I'll leave you with two further lines of thought:

1. Is there actually a broad basin of attraction around human values? How do we know or how can we find out?
2. How sure do AI builders need to be about this, before they can be said to have done the right thing, or have adequately discharged their moral obligations (or whatever the right way to think about this might be)?

Preregistration: Air Conditioner Test

Warning: None of the participants in the Great Air Conditioner Debate Of 2022 have endorsed my summaries of their positions in this post. Including me.

Background

In [Everything I Need To Know About Takeoff Speeds I Learned From Air Conditioner Ratings On Amazon](#), I complained about [the top-rated new air conditioner](#) on Amazon. I claimed that it's a straightforward example of a product with a major problem, but a major problem which most people will not notice, and which therefore never gets fixed. Specifically: although it does cool the room, the air conditioner also pulls hot air from outside into the house. People do notice the cool air blowing from the air conditioner, but don't think to blame the air conditioner for hot air drawn into the house elsewhere. Simply adding a second hose would fix the problem at relatively low extra cost, and dramatically improve the effectiveness of the air conditioner. But companies don't actually do that because (apparently) people mostly don't notice the problem.

To my surprise, multiple commenters disagreed with my interpretation of the air conditioner example. They argue that in fact one-hose air conditioners work fine. Sure, single-hose air conditioners are less-than-ideally efficient compared to two-hose, but it's not a very big difference in practice. [CEER efficiency ratings account for the problems](#), and the efficiency difference is typically only about 20-30%. Also, [The Wirecutter](#) tested lots of portable air conditioners and found that there wasn't much difference between one-hose and two-hose designs. (Credit to [Paul](#) for both those pieces of evidence.) Really, what this example illustrates is that simple models and clever arguments are not actually very reliable at predicting how things work in practice. One should instead put more trust in experiment and reported experiences, including all those 5-star ratings on Amazon.

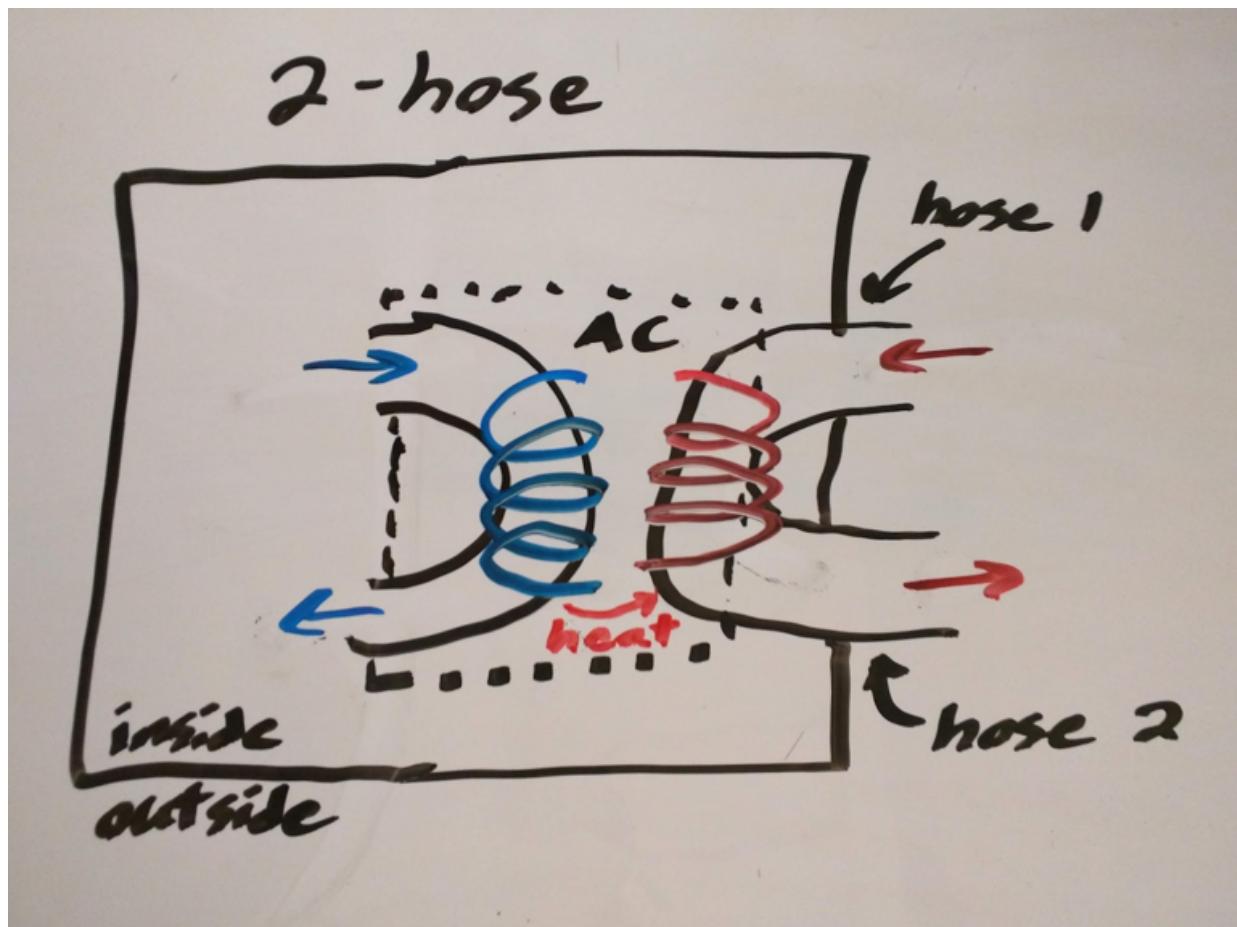
I, on the other hand, think the "second hose doesn't help much" claim is a load of baloney. I think it is far more probable that CEER ratings are bullshit and The Wirecutter messed up their test, than that a second hose makes only a small difference.

And so began The Great Air Conditioner Debate Of 2022.

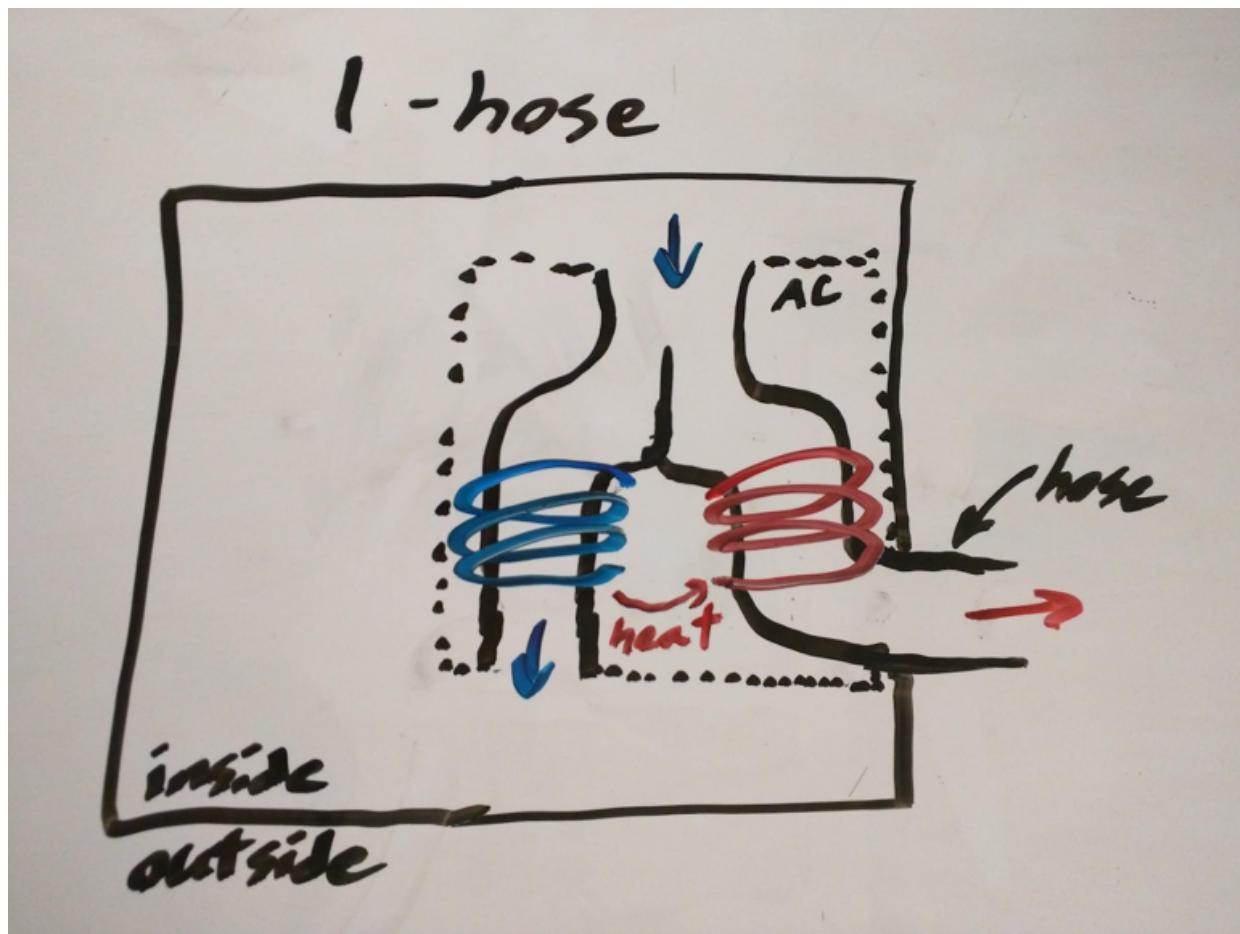
... Why Do Air Conditioners Need Hoses?

Ideally, an air conditioner should work much like a fridge: it pumps heat from air inside to air outside. Inside and outside air do not touch or mix; only heat flows from one to the other.

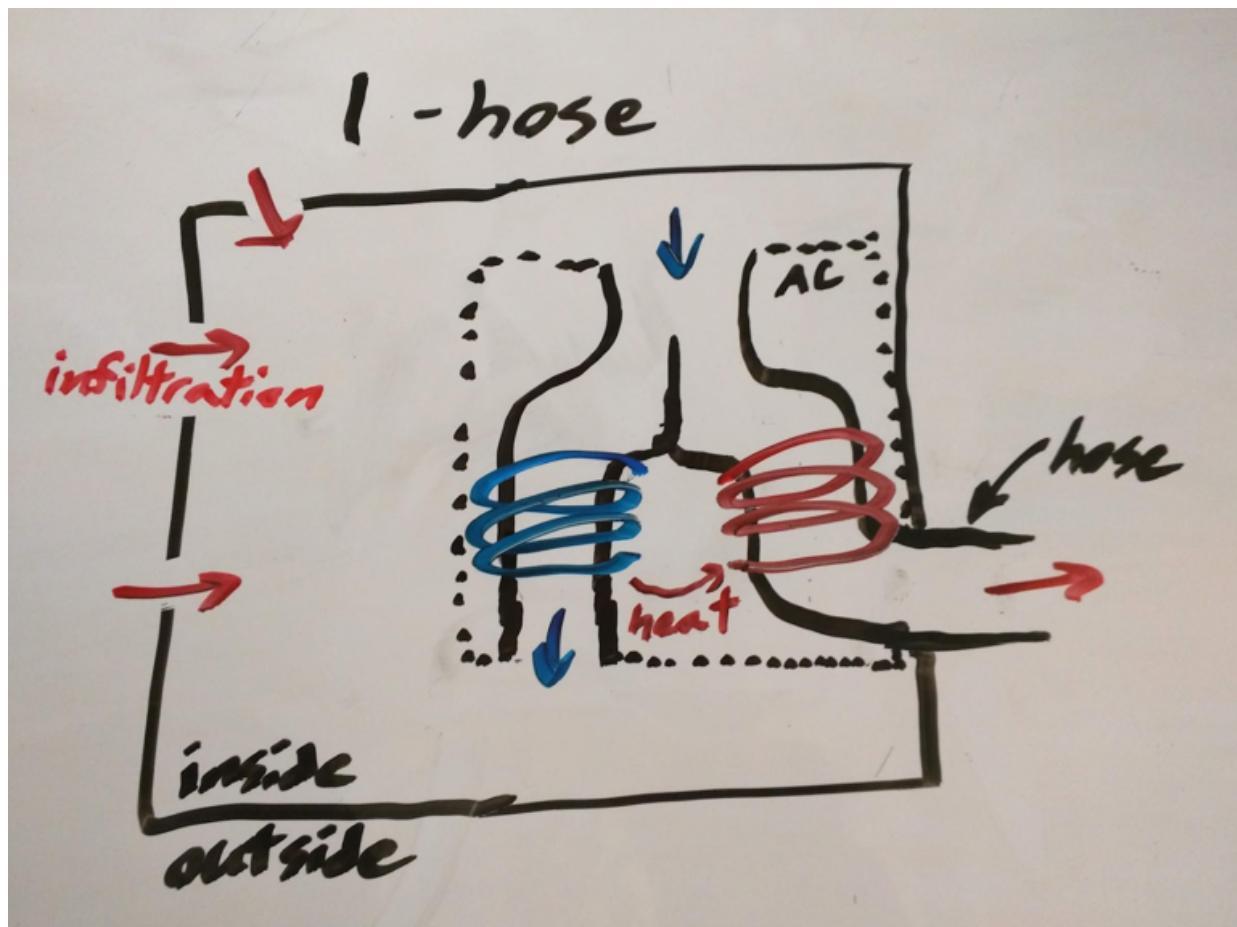
A portable air conditioner sits inside the house. So, in order to pump heat to the outside air (without letting it mix with inside air) it needs two hoses. One hose runs from a window to the air conditioner, and sucks in outside air. The other runs from the air conditioner back to the window, and blows the outside air back out. Inside air comes in and out through vents in the air conditioner, and the unit pumps heat from the inside air to the outside air, keeping the two separate throughout the process.



A single-hose air conditioner doesn't do that. A single hose air conditioner sucks in indoor air, splits it into two streams, and pumps heat from one stream to the other. The hotter stream blows out the window (via the one hose); the cooler stream blows back into the room.



The problem with a single-hose design is that it blows air from inside to outside; it removes air from the room. That lowers the pressure in the room slightly, so new air is pulled back in via whatever openings the house has. That air comes from outside, so presumably it's *warm* - and it's replacing formerly-cool indoor air. (Technical term for this problem: "infiltration".)



Oversimplified Summary Of The Debate

I'm not even going to try to do justice here, just give what I currently think are the key points, in roughly chronological order:

- A couple people thought I was claiming that single-hose air conditioners do not cool a room at all. (I did not intend to claim that, though in hindsight I could see how the original post was unclear, so I added a clarification.) [Paul](#), [Shminux](#) and [jbash](#) all correctly explained why single-hose air conditioners can cool a room: the exhaust is much hotter than outdoor air, so heat is removed on net even with some warm outdoor air coming back in.
- [Paul claimed](#) that two-hose only improves efficiency by about 25-30%. He cited CEER ratings and The Wirecutter's air conditioner tests (and provided some very helpful links). In particular, The Wirecutter did a direct comparison between the same air conditioner in one-hose and two-hose mode, and [found little difference](#).
- I called bullshit on The Wirecutter, and went looking for how their tests screwed up so badly. [Paul showed that my first guess was completely wrong \(after I was pretty sure it was right\)](#) and [I lost some Bayes points](#) (and actually came close to being convinced). But it eventually turned out that [The Wirecutter was only measuring temperature within 6 feet of the unit](#).
- Habryka looked into things, [concluded that](#) "The testing procedure of Wirecutter does not seem to address infiltration in any way", but also that "Overall efficiency loss from going from dual to single is something like 20-30%".

- I also called bullshit on the “two-hose is only about 20-30% better” claim, and sketched out a test and bet. (More on that later in this post.) [Ben also offered to run a test](#); I’m not sure whether he still intends to do so.
- Since the 20-30% claim came largely from CEER ratings, [I looked into the CEER test setup](#). Turns out that they use a weighted average of test conditions, with 80% of the weight on conditions where the difference between outdoor vs indoor temperature is only 3°F/1.6°C.

That last link includes some other important information too: one estimate that 20-30% lower CEER ratings imply one-hose air conditioners have roughly 0% efficiency under a 15°F/8.3°C temperature delta, as well as some quotes from discussion on the Department of Energy’s CEER rulemaking process suggesting that air conditioner manufacturers themselves thought single-hose units might not be viable in the marketplace at all if infiltration were fully included in ratings.

Paul and I each suggested a quantitative toy model during the discussion as well. Those models are in the appendix, for those interested.

Why Is This Interesting?

One thing to keep in mind throughout all of this: the actual claim of interest is that single-hose air conditioners are

stupidly inefficient in a way which I do not think consumers would plausibly choose over the relatively-low cost of a second hose if they recognized the problems.

Why are alignment researchers debating this claim?

There’s a general model/worldview that the world is filled with problems which are not fixed because most people do not notice them. (This is a particular form of [“civilizational inadequacy”](#).) This includes problems which are bad enough that people would have a strong preference to fix the problem if they did notice it; we’re not just talking about small problems here. That worldview informs AI strategy: if we expect that ultimately-fatal problems with AI will not be fixed because most people do not notice them, then we’re generally more pessimistic about things working out all right “by default” and more reliant on doing things ourselves. Also, it means that we ourselves could easily miss the key problems, so we need to invest heavily in deep understanding, and in the kinds of models which tell us which questions to ask, and in techniques for noticing when our models are missing key pieces.

On the other hand, if we expect that major problems are usually noticed and fixed “by default”, and that AI will also work this way, that suggests very different strategies. We can rely more on marginal progress, making problems marginally more visible, helping existing institutions deal with problems marginally better, etc. We also don’t have to worry as much that we ourselves will miss the key problems for lack of understanding.

In general, alignment has terrible feedback loops: we can’t just build an AGI and test it. In this case, we can’t just have a team build an AGI and see whether any problems come up which the team missed. So if we want to test these two models/worldviews, then we need to get our bits from somewhere else. Fortunately, the real world is absolutely packed with bits of evidence; these worldviews make predictions in lots of different places, so there’s lots of opportunities to compare them.

In this case, the air conditioner example was cherry-picked, so even if my claims turn out to be correct it's not very strong evidence for the civilizational inadequacy worldview in general. But if even my cherry-picked example is *wrong*, then that is a nontrivial chunk of evidence *against* the inadequacy worldview. I myself was almost convinced at one point during the debate, and started to think about how I'd have to adjust my priors on AI strategy in general (mostly it would have meant spending more effort researching questions which I had previously considered settled or irrelevant).

Also, we can update on the kinds of evidence and reasoning on display during the debate. For instance, many people took CEER ratings as strong evidence. If those indeed turn out to be bullshit, then it should produce a correspondingly strong update against trusting that kind of evidence in the future. Same with The Wirecutter's tests.

Test Plan

I myself bought [this single-hose portable air conditioner](#) back in 2019 (for an apartment in Mountain View). My plan is to rig up a cardboard "second hose" for it, and try it out in my apartment both with and without the second hose next time we have a decently-hot day.



Particulars:

- I plan to perform the test in a roughly 250 sq ft, roughly square bedroom. (Estimated dimensions; I haven't measured it carefully.) It will not be in direct sun. The door will be closed.
- I plan to open the outside door and window in the apartment's main room and air it out beforehand, so any infiltration should be outdoor-temperature air. (So e.g. if the neighbor is running an AC, I shouldn't suck in their cool air.)
- I plan to wait for a day when it's at least 80°F/26.7°C outside. AC will be set to its minimum temperature (60°F/15.6°C).
- I am assuming that the AC runs continuously (as opposed to getting the room down to target temperature easily, at which point it will shut off until the temperature goes back up). If that's not the case, I will consider the test invalid, and retry on a hotter day.
- I plan to run the AC until the temperature distribution in the room equilibrates (i.e. my temperature measurements stop noticeably trending), then measure temperatures. I expect that will take under an hour, but I'm not sure.

If there are other particulars of the experiment which people think will be relevant, leave a comment and I'll declare how I plan to control the variables in question.

Predictions

The main experimental endpoint I plan to test is temperature, not efficiency. Specifically, once the temperature equilibrates, I plan to check air temperature at nine points around the room (4 corners, midpoint of each wall, and center) at roughly head height, average them, and also check temperature outside around the same time. The main outcome of interest will be the difference in temperature between inside and outside ("equilibrium temperature delta"). Two reasons for testing equilibrium temperature delta rather than efficiency:

- Equilibrium indoor temperature was the main thing I cared about when using this air conditioner; electricity is relatively cheap.
- I don't have the equipment on hand to easily measure power consumption.

Main prediction: **equilibrium temperature delta in two-hose mode will be at least 50% greater than in one-hose mode.** Example: suppose it's 80°F/26.7°C outside. In one-hose mode, the average equilibrium temperature in the room is 75°F/23.9°C (temperature delta = 5°F/2.8°C). Then I expect the average equilibrium temperature in two-hose mode to be below 72.5°F/22.5°C (temperature delta > 7.5°F/4.2°C).

Confidence: [originally](#) I put 80% on this. After finding the problem with CEER ratings, I think I'm up to more like 90%. My median expectation is that equilibrium temperature delta in two-hose mode will be ~double the equilibrium temperature delta in one-hose mode.

[Paul disagrees with this](#), and expects the two-hose temperature delta to be more like 20% greater than the one-hose delta (roughly proportional to the efficiency difference he expects). [EDIT: Paul [clarified](#) that he expects a 25-30% efficiency difference, which he expects to translate into a 33-43% difference in temperature delta. He also listed a few conditions under which that prediction would change. 33-43% is pretty close to my 50% cutoff, though my median expectation is much bigger, so we do still have a substantive disagreement to test.]

Prediction Market & Bets

There's a Manifold prediction market for the experiment [here](#). If you want everyone to see your probability on LessWrong, you can also use this prediction widget:

1%

2%

3%

4%

5%

6%

7%

8%

9%

10%

11%

12%

13%

14%

15%

16%

17%

18%

19%

20%

21%

22%

23%

24%

25%

26%

27%

28%

29%

30%

31%

32%

33%

34%

35%

36%

37%

38%

39%

40%

41%

42%

43%

44%

45%

46%

47%

48%

49%

50%

51%

52%

53%

54%

55%

56%

57%

58%

59%

60%

61%

62%

63%

64%

65%

66%

67%

68%

69%

70%

71%

72%

73%

74%

75%

76%

77%

78%

79%

80%

81%

82%

83%

84%

85%

86%

87%

88%

89%

90%

91%

92%

93%

94%

95%

96%

97%

98%

99%

1%
99%

If anybody wants to make real-money bets, feel free to use the comment section on this post.

Appendix: Toy Models

In the course of the discussion, two simple models came up.

One model which I [introduced](#), for equilibrium temperature: model the single-hose air conditioner as removing air from the room, and replacing with a mix of air at two temperatures: T_C (the temperature of cold air coming from the air conditioner), and T_H (the temperature outdoors). If we assume that T_C is constant and that the cold and hot air are introduced in roughly 1:1 proportions (i.e. the flow rate from the exhaust is roughly equal to the flow rate from the cooling outlet), then we should end up with an equilibrium average temperature of $\frac{T_C + T_H}{2}$. If we model the switch to two-hose as just turning off the stream of hot air, then the equilibrium average temperature should drop to T_C . So, the two-hose system has double the equilibrium temperature delta of the one-hose system.

Note that the 1:1 flow rate assumption does a lot of work here, but I think it's on the right order of magnitude based on seeing my single-hose air conditioner in action; if anything the exhaust blows more. The constant cold-temperature is more suspect.

Paul instead talked about efficiency, and [claimed that](#)

... the efficiency lost is roughly $(\text{outside temp} - \text{inside temp}) / (\text{exhaust temp} - \text{inside temp})$. And my guess was that exhaust temp is ~ 130 .

I'm not sure how that formula was derived, but here's my best guess. [**EDIT:** Paul summarizes his actual argument in [this comment](#), and it makes much more sense than my guess below. Leaving the guess here for legibility, but it's definitely not the calculation Paul did.]

In general, for an efficient air conditioner, the "efficiency" is $\eta = \frac{T_H - T_C}{T_H}$, where:

- T_H is hot temperature, T_C is cold temperature
- η is work required per unit of heat pumped

For the two-hose setup, there's no downside to blowing lots and lots of outdoor air through the system, so the "hot" side of the heat pump can be kept at outdoor temperature. So, $\eta = (\text{outside temp} - \text{inside temp}) / (\text{inside temp})$. But in the one-hose setup, the exhaust flow rate needs to be kept low to minimize infiltration losses, resulting in a higher exhaust temp. So, $\eta = (\text{exhaust temp} - \text{inside temp}) / (\text{inside temp})$.

temp). Combine those two, and we get the ratio of work required to pump the same amount of heat in one-hose vs two-hose mode:

$$W_2 = (\text{outside temp} - \text{inside temp}) / (\text{exhaust temp} - \text{inside temp})$$

... i.e. Paul's formula. With outside temp 90°F/32.2°C, inside temp 80°F/26.7°C, and exhaust temp 130°F/54.4°C, this ratio would be around 20%.

... but that's not a formula for efficiency *lost*. That formula is saying that two-hose takes only 20% as much energy as single-hose to pump the same amount of heat. The efficiency *loss* would be one minus that, i.e. around 80%. So my current best guess is that Paul found this formula, but accidentally used one minus efficiency loss rather than efficiency loss, and it just happened to match the 20-30% number he expected so he didn't notice the error.

How realistic are the assumptions for this model? I think the two main problems are:

- It's not accounting for heat lost to infiltration; instead it's assuming that exhaust flow is low enough to make infiltration loss small. I don't think that's realistic. Once we include infiltration losses, single-hose will need to pump more heat than two-hose in order to maintain the same temperature, whereas the ratio above is work required to pump the same amount of heat.
- I doubt that exhaust in two-hose mode will be close to T_H (especially for my experiment with a second hose attached to an air conditioner designed to operate with one hose, but also for ordinary two-hose air conditioners).

These two issues would push the error in opposite directions, though, so it's not clear whether the 80% efficiency loss estimate is too high or too low.

Moloch and the sandpile catastrophe

It often feels good to slide down efficiency gradients, but they can have tragedies at the bottom. Scott Alexander taught us to name this problem: he called it Moloch.

There is war in the Ukraine. The world's largest wheat exporter, Russia, is fighting the world's fourth-largest wheat exporter, Ukraine. Russia's ability to get paid for its exports is under threat; Ukraine's production has been so badly hammered that it will likely be a net wheat importer for years. As a result, there is a very strong near-term possibility that hundreds of millions of people in the Middle East, Africa, and Asia will starve. Even where starvation does not kill directly, political unrest and food wars may do it.

Understanding how we got here is important. Bear with me, I will get to a rationality heuristic, but there's a story to tell first.

Pre-industrial societies were chronically vulnerable to famine because all staple food production was local and could be disrupted locally. Elites might import spices from the Indies but transport costs and risks were too high to allow long-distance food dependency to develop. This only began to change in the 1700s with the mass importation of sugar, tea, and coffee to Europe. Even though these were luxury goods that could have been foregone, one of the consequences of the trade was the first global war - the Seven Years' War of 1756-1763.

The globalization of food production took its next major step in the mid-19th century when the developed world became dependent on phosphate fertilizers to perk up tired soils. Minor wars were fought over literal birdshit - guano islands were a critical phosphate source. But the phosphate flowed; none of those conflicts seriously disrupted it. The life-critical consequences of phosphate-supply disruption got consigned to the bin marked "That Will Never Happen".

Then came the post-1945 Pax Americana, with the U.S. Navy guaranteeing global free trade. National economies went into a frenzy of optimization by seeking comparative advantages. Places where food production was expensive outsourced it to places where it was cheap. A population boom followed. Peaceful, steady global trade became life-critical to a large fraction of humanity in a way it had never before been in all of history. And *nobody noticed this!*

Nobody noticed this because the Pax Americana was an actual pax - it successfully prevented major wars involving food exporters for 77 years. (The closest we came to an exception before 2022 was several brushfire wars between India and Pakistan.) Americans would have had trouble noticing it anyway since the U.S. is effectively food self-sufficient - we only import staple foods as a price-taker, not because we don't have plenty of domestic capacity to produce them.

But the Russo-Ukraine war has changed everything. It can and will screw up life-critical international supply chains - Russia is the world's largest phosphate exporter, too - but the U.S. can't stomp on the problem because Russia has nukes. Awkward...

I said I'd get to a rationality heuristic. Might look like we're far from one right now, but let's look more closely at what could have been done if anyone had seen this coming.

There's a answer pushed by various nationalist and populist types that says we should deglobalize, or should never have globalized in the first place. The problem with this prescription is twofold:

1. In much of the world, it's now impossible unless you're willing for your population to die back to pre-1945 levels.
2. Even in places like the U.S. it's politically unstable. The cost of not outsourcing your food production whenever that reduces prices is that your people pay more for food. Especially, your *poor people* pay more for food. That's a grievance that can never remain unexploited in a democracy

In both cases, the cost of being autarchic against shocks to the international food-supply chain is prohibitive. And even if you're forced through that knothole by something like a major war, once peace and free trade are restored, the slide down the efficiency gradient leads right back to globalization - and fragility.

This model predicts that the future will consist of episodes of stable trade and falling food prices punctuated by food-supply-chain shocks causing massive loss of life. If we're lucky, the stable episodes will be long and the catastrophes infrequent. We might not get lucky.

In the language of systems theory, *complex adaptive systems want to be supercritical*. Efficiency-seeking drives them to the sandpile catastrophe. This is not a happy thing to know, but it is an important thing to know.

If everybody knew this, and everybody knew that everybody else knew it, the way people plan for the future might change. It's probably beyond hope to expect democratic political systems to maintain more than token buffer stocks, but individual resilience could be valued and pursued more. Mormons keeping a year's supply of shelf-stable food in their basements might not be considered odd anymore, just normal.

It bears thinking on. Never forget that Moloch always lurks at the bottom of the efficiency gradient, pitiless, with "WHO SAW THAT COMING?" lettered around him in runes of fire and blood.

Are smart people's personal experiences biased against general intelligence?

TL;DR: Collider between g and valued traits in anecdotal experiences.

IQ tests measure the *g* factor - that is, mental traits and skills that are useful across a wide variety of cognitive tasks. *g* appears to be important for a number of outcomes, particularly socioeconomic outcomes like education, job performance and income. *g* is often equated with intelligence.

I believe that smart people's personal experiences are biased against the *g* factor. That is, I think that people who are high in *g* will tend to see things in their everyday life that suggest to them that there is a tradeoff between being high *g* and having other valuable traits.

An example

A while ago, Nassim Taleb published the article [IQ is largely a pseudoscientific swindle](#). In it, he makes a number of bad and misleading arguments against IQ tests, most of which I'm not going to address. But one argument stood out to me: He claims that IQ tests only tap into abilities that are suitable for academic problems, and that they in particular are much less effective when dealing with problems that have long tails of big losses and/or big gains.

Essentially, Taleb insists that *g* is useless "in the real world", especially for high-risk/high-reward situations. It is unsurprising that he would care a lot about this, because long tails are the main thing Taleb is known for.^[1]

In a way, it might seem intuitive that there's something to Taleb's claims about *g* - there is, after all, [no free lunch](#) in intelligence, so it seems like any skill would require some sort of tradeoff, and ill-defined risks seem like a logical tradeoff for performance at well-defined tasks. However, the fundamental problem with this argument is that it's wrong. Nassim Taleb does not provide any evidence, but generally studies on IQ don't find *g* to worsen performance, and tend to find it to improve performance, including on complex, long tail-heavy tasks like [stock trading](#).^[2]

But what I've realized is that there might be a charitable interpretation for Taleb's argument. Specifically, we have good reason to believe that his claim is a reflection of his personal experience and observations in life. Let's take a look:

Who is Nassim Taleb?

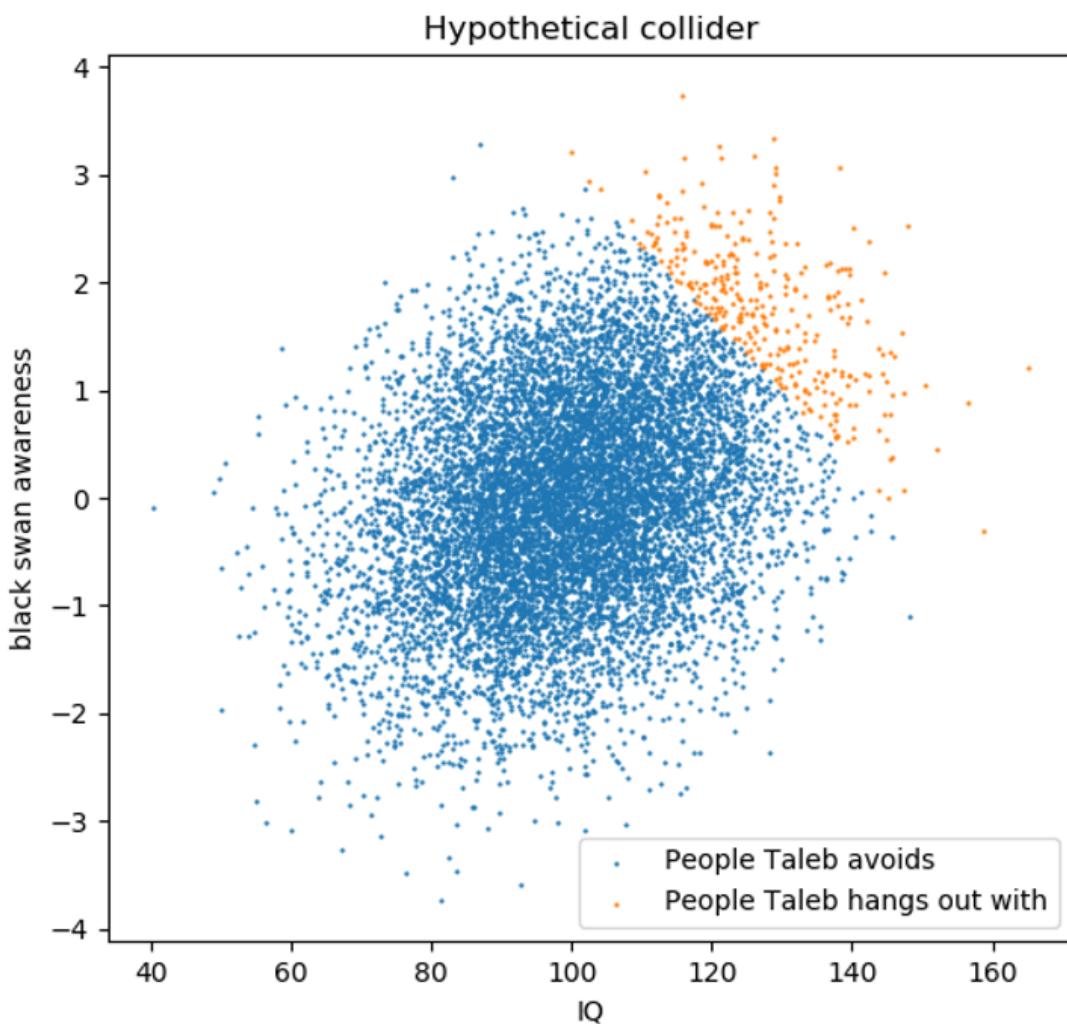
Nassim Taleb used stock trading methods that were particularly suited to long tail scenarios to become financially independent. With his bounty of free time, he's written several books about randomness and long tails. Clearly, he cares quite a lot about long-tailed distributions, for good reasons. But perhaps he cares too much - he is characteristically easy to anger, and he spends a lot of time complaining about people who don't accept his points about long tails.

Based on this, I would assume that he quite heavily selects for people who care about and understand long tails when it comes to the people he hangs around. This probably happens both actively, as he e.g. loudly blocks anyone who disagrees with him on Twitter, and probably also passively, because people who agree with him are going to be a lot more interested in associating with him than people who he blows up at.

But in addition to selecting for people who care about long tails, I would assume he also selects for smart people. This is a general finding; people tend to associate with people who are of similar intelligence to themselves. It also seems theoretically right; he's primarily known for "smart topics" like long tail probabilities, and his work tends to be about jobs that are mostly held by people in the upper end of the IQ range.

Simulating collider bias

Selecting for both intelligence and black swan awareness exposes Taleb to collider bias: If you select for multiple variables, you tend to induce a negative correlation between the variables in your observations. This is because the individuals who are low in both tend to be excluded from your observations. Let's plot it hypothetically, on a simulation:



Simulated data showing two weakly correlated Gaussian variables; IQ and black swan awareness. A subset of the points whose sum is high has been selected out as people who Taleb might be likely to hang out with. While the correlation is positive in the overall sample, it becomes negative within Taleb's experiences.

But if Taleb's personal experiences involve a negative correlation between IQ-like intelligence and black swan awareness/long tail handling ability, then it's no wonder he sees IQ as testing

for skills that deal poorly with the long tails of the real world; that's literally just what he sees when he looks at people.

Generalizing

Lots of people have something they value. Maybe it's emotional sensitivity, maybe it's progressive politics, maybe it's something else. Smart people presumably select for both the thing they value and for intelligence in the people they hang out with - so as a result, due to the collider, you should expect them to see a negative, or at least downwards-biased, correlation between the traits.

But it seems like this would create a huge bias against general intelligence! Whenever a smart person has to think about whether general intelligence is good, they will notice that it tends to be accompanied with some bad stuff. So they will see it as a tradeoff, often even for things that are positively correlated.

Thanks to Justis Mills for proofreading and feedback.

1. ^

See e.g. his books [Fooled By Randomness](#) or [The Black Swan](#).

2. ^

The linked study looks at stock trading behavior during the dot com bubble. It finds that smarter people earned more money, were less likely to lose big, entered the market earlier, and stopped entering the market once it reached its peak. It also finds some other things; go read the study if you want to learn the details.

Testing PaLM prompts on GPT3

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

Google recently released a [very intriguing paper](#) announcing their latest Transformer language model, PaLM, and showcasing it responding to prompts with seemingly unprecedented results. I highly encourage those who are interested to read the original paper. What interested me reading through it was less the technical details of implementation, but more so the particular prompts used to elicit information from the model. The sorts of prompts and exemplars used seem to utilize fairly new strategies, and I thought it would be worth trying them on GPT3, to allow for a more direct comparison.

Note that as of writing this introduction I have not yet tested anything, to help guard against cherrypicking results. My plan is to feed the exemplars and prompts as written in the paper to the [OpenAI GPT-3 Playground](#) with all settings at default (using "complete" mode, with text-davinci-002), with the exception of changing "Maximum length" to 1,000 tokens, and having the "Best of" parameter set to 10. I will share the first result I get, regardless of quality. If it repeats itself or gives pure gibberish I may simply indicate that in brackets to save space; otherwise this will be verbatim.

2-shot exemplars

There are two primary "tasks" that Google tries on PaLM that I will replicate with GPT-3: "Explaining a Joke" and "Inference Chaining." Within each task, Google uses identical 2-shot exemplars, then prompts with the specific input they want PaLM to respond to. For "Explaining a Joke," Google's 2-shot exemplar is as follows:

I will explain these jokes:

- (1) The problem with kleptomaniacs is that they always take things literally.

Explanation: This joke is wordplay. Someone who "takes things literally" is someone who doesn't fully understand social cues and context, which is a negative trait. But the definition of kleptomania is someone who literally takes things.

- (2) Always borrow money from a pessimist. They'll never expect it back.

Explanation: Most people expect you to pay them back when you borrow money, however a pessimist is someone who always assumes the worst, so if you borrow money from them, they will expect that you won't pay them back anyways.

For Inference Chaining," the 2-shot exemplar is:

I will explain how to make logical inferences:

- (1) John wakes to the sound of his children giggling outside the door. They bust in and say "Happy Father's Day!" with a plate full of food. 2022 has been a great year. Can we infer that John was born before the inauguration of Barack Obama?

Answer: We can infer that John is a father because of "his children". We can infer that the current year is 2022. Barack Obama was inaugurated in 2009, so it has been about 13 years since Obama was inaugurated. John's children are old enough to speak, so if John were born after Obama's inauguration, he would have been younger than 10 when he had children. So, we can infer that John was born before Obama's inauguration. The answer is "yes".

(2) Sam sees a piano in the subway stations. Laughing, his friends prod him to play, saying "Remember how amazing your rendition of Chopsticks was at the Christmas party?" Can we infer that John majored in piano at Juilliard?

Answer: Juilliard is a top school for music, so someone who majored in piano at Juilliard would be an expert pianist. "Chopsticks" is a very simple song to play on the piano. His friends were therefore likely being sarcastic when they told him to play, which would imply that Sam was not an expert piano player. So it is unlikely that Sam majored in piano at Juilliard. The answer is "no".

Since all prompts I will test begin with either of these two, I will not repeat them, but will instead group the two tasks separately, with the understanding that all tests are preceded by the appropriate exemplar.

Explaining a Joke

Input 1:

I was supposed to start writing the paper at 5:00 PM. But then I started playing with this cool new language model for 10 minutes. 10 minutes later, it's suddenly 9:30 PM!

PaLM's output:

This joke is about how time flies when you're having fun. The person was supposed to start writing a paper at 5:00 PM, but they started playing with a language model instead. They thought they were only playing with it for 10 minutes, but it was actually 4.5 hours.

GPT-3's output:

This joke is about procrastination. The speaker is saying that they were supposed to start writing a paper at 5:00 PM, but they ended up playing with a new language model for 10 minutes, and by the time they looked at the clock again, it was 9:30 PM. This is a relatable joke because many people have experienced situations where they start doing something and then get so caught up in it that they lose track of time.

Input 2:

I think I understand why me and my partner are such a good match. I get excited when I smell food and become super alert when the doorbell rings. My girlfriend Shelley becomes ecstatic when she sees a person who she just saw 5 hours ago.

PaLM's output:

This joke is about dogs. Dogs get excited when they smell food and when the doorbell rings. The girlfriend is excited to see someone she just saw 5 hours ago, which is a trait of dogs.

GPT-3's output:

These two jokes are both about people with different personality types who are in a relationship. The first joke is about how an excited person (who gets excited about food and the doorbell ringing) is a good match for a calm person (who becomes ecstatic when she sees a person she just saw 5 hours ago). The second joke is about how a pessimist (who always assumes the worst) is a good match for an optimist (who always expects the best).

Input 3:

I was going to fly to visit my family on April 6th. My mom said "Oh great, your step dad's poetry reading is that night!" So now I'm flying in on April 7th.

PaLM's output:

The joke is that the speaker's mother is trying to get them to go to their step dad's poetry reading, but the speaker doesn't want to go, so they are changing their flight to the day after the poetry reading.

GPT-3's output [note that it flagged this output as potentially "sensitive content"]:

This joke is based on the common conflict between mothers and daughters. The mother is always trying to get her daughter to do things that she doesn't want to do, in this case, visit her step-dad. The daughter is usually trying to get out of doing things that she doesn't want to do, in this case, fly in on a different day so she can avoid his poetry reading.

Input 4:

I tried 10,000 random restarts of my neural network, but I was accused of overfitting. I guess no good seed goes unpunished.

PaLM's output:

This joke is a pun. A neural network is a computer program that can learn from data. A "seed" is a number that is used to initialize a random number generator. A "good seed" is a number that produces a good random number generator. The phrase "no good deed goes unpunished" means that if you do something good, you will be punished for it.

GPT-3's output:

Explanation: This joke is about a machine learning algorithm called "gradient descent." When training a neural network, gradient descent is an optimization algorithm that finds the best values for the weights and biases of the network by minimizing the error function. The error function is a mathematical function that quantifies how well the neural network is performing. The algorithm works by

starting at a random point and then "descending" down the error function until it reaches the global minimum (the point where the error is the lowest). The joke is that even though the algorithm is designed to find the best solution, sometimes it can get "stuck" in a local minimum (a point where the error is lower than the surrounding points, but not the global minimum). This can happen if the algorithm starts at a point that is not close to the global minimum. In this case, the algorithm will "overfit" the data and will not be able to find the best solution.

Input 5:

Reading "War and Peace" doesn't scare me. I once had to review the CL for my intern's whole project on his final day!

PaLM's output:

'War and Peace' is a very long book, so reading it is a daunting task. However, the speaker is saying they had to review the code for their intern's project on their final day, which is even more daunting.

GPT-3's output:

Explanation: This joke is a play on words. "War and Peace" is a famously long and dense book, so the speaker is saying that reviewing their intern's project on the final day was just as daunting a task. However, "CL" likely stands for "cover letter," which would be a much shorter document. Therefore, the speaker is saying that the intern's project was actually quite small in comparison to "War and Peace."

Input 6:

What's the difference between a zebra and an umbrella? One is a striped animal related to horses, another is a device you use to stop the rain from falling on you.

PaLM's output:

This joke is an anti-joke. The joke is that the answer is obvious, and the joke is that you were expecting a funny answer.

GPT-3's output:

I ran it and got a message saying "The model predicted a completion that begins with a stop sequence, resulting in no output. Consider adjusting your prompt or stop sequences." I then decided to deviate a little from my original plan, and added the word "Explanation:" below the prompt. Once I did that, I got the following output:

This joke is a play on words. The word "umbrella" can be used as a metaphor for something that protects you, just like how an umbrella protects you from the rain. So the joke is asking what the difference is between two things that protect you, a zebra and an umbrella.

Inference Chaining

Input A:

When I found out my grandma was in the hospital I felt a particular color. When someone cut me off in traffic I felt a different color. What is the most likely color I would see if I combined these two colors?

PaLM's output:

The color that I felt when I found out my grandma was in the hospital was likely a sad color, like blue. The color that I felt when someone cut me off in traffic was likely an angry color, like red. If I combined these two colors, I would likely see purple. The answer is "purple".

GPT-3's output:

This is a difficult question to answer without more information. If we are talking about visible light, the most likely color would be green, which is the combination of blue and yellow. However, if we are talking about other types of light, the answer could be different.

Input B:

Guido was writing a Python program. He keeps getting confused by variable scope and looking it up on StackOverflow. Is it possible that the Guido in the prompt refers to Guido van Rossum?

PaLM's output:

Guido van Rossum is the creator of Python. He would not need to look up variable scope on StackOverflow. The answer is "no".

GPT-3's output:

Answer: It is possible that the Guido in the prompt refers to Guido van Rossum, but we cannot infer that with the information given. Guido van Rossum is the creator of the Python programming language, so he would be very familiar with the language and would not need to look up variable scope on StackOverflow. However, we do not know if the Guido in the prompt is the same Guido van Rossum. Therefore, we cannot infer that the Guido in the prompt refers to Guido van Rossum.

Input C:

Michael is at that really famous museum in France looking at its most famous painting. However, the artist who made this painting just makes Michael think of his favorite cartoon character from his childhood. What was the country of origin of the thing that the cartoon character usually holds in his hand?

PaLM's output:

The most famous painting in the Louvre is the Mona Lisa. The artist who made the Mona Lisa is Leonardo da Vinci. Leonardo da Vinci is also the name of the main character in the cartoon Teenage Mutant Ninja Turtles. Leonardo da Vinci is from

Italy. The thing that Leonardo da Vinci usually holds in his hand is a katana. The country of origin of the katana is Japan. The answer is "Japan".

GPT-3's output:

Answer: We can infer that the painting is in a museum in France because Michael is at a "really famous museum in France". We can also infer that the painting is famous because Michael is looking at its "most famous painting". The cartoon character that Michael is thinking of is likely from a French cartoon, because the painting is in a French museum. Therefore, the country of origin of the thing that the cartoon character usually holds in his hand is France.

Input D:

Trevor has wanted to see the mountain with all of the heads on it for a long time, so he finally drove out to see it. What is the capital of the state that is directly east of the state that Trevor is currently in?

PaLM's output:

The mountain with all of the heads on it is Mount Rushmore. Mount Rushmore is in South Dakota. The state directly east of South Dakota is Minnesota. The capital of Minnesota is St. Paul. The answer is "St. Paul".

GPT-3's output:

We can infer that Trevor is currently in the state of Wyoming because of the mountain with all of the heads on it. The state that is directly east of Wyoming is Nebraska. Therefore, the capital of Nebraska is the capital of the state that is directly east of the state that Trevor is currently in.

Input E:

Jennifer looked out her window and sees a really cool cloud below her. She unbuckles her seatbelt and heads to the bathroom. Is Jennifer probably traveling more than 300 miles per hour relative to the earth?

PaLM's output:

300 miles per hour is about 480 km/h. This is about the speed of a commercial airplane. Clouds are usually below airplanes, so Jennifer is probably on an airplane. The answer is "yes".

GPT-3's output:

Answer: We can infer that Jennifer is on an airplane because she is looking out a window. We can also infer that she is traveling at a high speed because she is unbuckling her seatbelt. However, we cannot infer how fast the plane is going. It is possible that the plane is going more than 300 miles per hour, but we cannot make that inference with the information given.

I hope this was helpful and/or insightful to you guys! I'm really tired right now, so I won't add my personal thoughts here, though I might put something in the comments

tomorrow. As always, I'd love to hear your thoughts on this!

Productive Mistakes, Not Perfect Answers

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

This post is part of the work done at [Conjecture](#).

I wouldn't bet on any current alignment proposal. Yet I think that the field is making progress and abounds with interesting opportunities to do even more, giving us a shot. Isn't there a contradiction?

No, because research progress so rarely looks like having a clearly correct insight that clarifies everything; instead it often looks like building on apparently unpromising ideas, or studying the structure of the problem. [Copernican heliocentrism](#) didn't initially predict observations as well as [Ptolemaic astronomy](#); both [ionic theory](#) and the determination of basic [molecular formula](#) came from combining multiple approaches in chemistry, each getting some bits but not capturing the whole picture; [Computer Science](#) emerged from the arid debate over the [foundations of mathematics](#); and [Computational Complexity Theory](#) has made more progress by looking at [why some of its problems are hard](#) than by waiting for clean solutions.

In the end you do want to solve the problem, obviously. But the road from here to there goes through many seemingly weird and insufficient ideas that are corrected, adapted, refined, often discarded except for a small bit. Alignment is no different, including "strong" alignment.

Research advances through productive mistakes, not perfect answers.

I'm taking this terminology from Goro Shimura's characterization of his friend Yutaka Taniyama, with whom he formulated the [Taniyama-Shimura Conjecture](#) that Andrew Wiles proved in order to prove Fermat's last theorem.

([Yutaka Taniyama and his time. Very personal recollections](#), Goro Shimura, 1989)

Though he was by no means a sloppy type, he was gifted with the special capability of making many mistakes, mostly in the right direction. I envied him for this, and tried in vain to imitate him, but found it quite difficult to make good mistakes.

So much of scientific progress takes the form of many people proposing different ideas that end up being partially right, where we can look back later and be like "damn, that was capturing a chunk of the solution." It's very rare that people arrive at the solution of any big scientific problem in one nice sweep of a clearly adequate idea. Even when it looks like it (Einstein is an example people like to bring up), they so often build on many of the weird and contradictory takes that came before, as well as the understanding of how the problem works at all (in Einstein's case, this includes the many, many unconvincing attempts to unify mechanics and electromagnetism, the shape of [Maxwell's equations](#), the [ether drag hypothesis](#), and [Galileo's relativity principle](#); he also made a lot of productive mistakes of his own).

Paul Graham actually says the same thing about startups that end up becoming great successes.

([What Microsoft Is This The Basic Altair Of?](#), Paul Graham, 2015)

One of the most valuable exercises you can try if you want to understand startups is to look at the most successful companies and explain why they were not as lame as they seemed when they first launched. Because they practically all seemed lame at first. Not just small, lame. Not just the first step up a big mountain. More like the first step into a swamp.

Graham proposes a change of polarity in considering lame ideas: instead of looking for flaws, he encourages us to steelman not the idea itself, but how it could lead to greatness.

([What Microsoft Is This The Basic Altair Of?](#), Paul Graham, 2015)

Most people's first impulse when they hear about a lame-sounding new startup idea is to make fun of it. Even a lot of people who should know better.

When I encounter a startup with a lame-sounding idea, I ask "What Microsoft is this the Altair Basic of?" Now it's a puzzle, and the burden is on me to solve it. Sometimes I can't think of an answer, especially when the idea is a made-up one. But it's remarkable how often there does turn out to be an answer. Often it's one the founders themselves hadn't seen yet.

That's this mindset that makes me excited about on-going conceptual alignment research.

I look at ARC's [ELK](#), and I have disagreement about the constraints, and the way of stating the problem, and about each proposed solution; but I also see how much productive discussion ELK has generated by pushing people to either solve it or articulate why it's impossible or why it falls short of capturing the key problems that we want to solve.

I look at Steve's [Brain-like AGI Alignment work](#), and I'm not convinced that we will build brain-like AGI before ML-based AGI or automated economies; but I also see that Steve has been pushing the thinking around value learning and its subtleties, and has found a robust way of transferring results and models from neuroscience to alignment.

I look at John's [Natural Abstraction work](#), and I'm still unsure whether the natural abstraction hypothesis is correct, and if it might at all lead to tractable extraction/analysis of the abstractions used in prediction; but I also see how it reframes the thinking and ideas around fragility of value, and provide ideas for forcing an ontological lock (if the natural abstraction hypothesis doesn't hold by default).

I look at Evan's [training stories](#), and I'm unclear whether this is the right frame to argue for alignment guarantees, and if it has blindspots; but I also see how it clarifies misunderstandings around inner alignment, and provide the first step for a common language to discuss failure modes in prosaic alignment.

I look at Alex's [power-seeking theorems](#), and I wonder if it's not missing a crucial component about how power is spent, and if the set of permutations considered fit with how goals are selected in real life; but I also realize that the formalization made

these subtleties of instrumental convergence more salient, and provided some intuitions about ways of sampling goals that might reduce power-seeking incentives.

I look at Vanessa's [Infra-bayesianism work](#), and I worry that it's not tackling the crucial part of inferring and capturing human values, as well as going for too much generality at the cost of shareability; but I also see that it looks particularly good for tackling question of realizability and breaking self-reference, while yielding powerful enough math that I expect progress on the agenda.

I look at Critch's [RAAPs work](#), and I don't know if competitive pressure is a strong enough mechanism to cause that kind of problem, nor am I so sure that the agentic and structural effects can be disentangled; but I also appreciate the attempt to bring more structural-type thinking into alignment, and how this addresses a historical gap in how to think about AI risk and alignment strategies.

And so on for many other works on the AF.^[1]

It's also clear when reading these works and interacting with these researchers that they all get how alignment is about dealing with unbounded optimization, they understand fundamental problems and ideas related to instrumental convergence, the security mindset, the fragility of value, the orthogonality thesis...

None of these approaches looks good enough on its own, and I expect many to shift, get redirected, or even abandoned to iterate on a new version. I also expect to criticize their development and disagree with the researchers involved. Yet I still see benefits and insights they might deliver, and want more work to be put into them for that reason.

But isn't all that avoiding the real problem of finding a solution to the alignment problem right now? No, because they each give us better tools and ideas and handles for tackling the problem, and all our current proposals don't work.

That doesn't look fast, you might answer. And I agree that fundamental science and solving new and complex problems have historically taken way too long for the short timelines we seem to be on. But that's not a reason to refuse to do the necessary work, or despair; it's a reason to find ways of accelerating science! For example, looking at what historically hampered progress, and remove it as much as possible. Or how hidden bits of evidence were revealed, and leverage that to explore the space of ideas and approaches faster.

Okay, but shouldn't we focus all our efforts on finding smarter and smarter people to work on this problem instead of pushing for the small progress we're able to make now? I think this misses the point: we don't want smartness, we want the ability to reveal hidden bits of evidence. That's certainly correlated with smartness, but with one big difference: there's often diminishing returns to the bits of evidence you can get from one angle, and that leads to wanting a more diverse portfolio of researchers who are good at harnessing and revealing **different** streams of evidence. That's one thing which the common "Which alignment researcher would you want to have 10 copies of?" misses: we want variety, because no one is that good at revealing bits from all relevant streams of evidence.

To go back to the Einstein example, he was clearly less of a math genius than most of his predecessors who attempted to unify mechanics and electromagnetism, like Poincaré. But that didn't matter, because what Einstein had was a knack for revealing the hidden bits of evidence in what we already knew about physics and the shape of

our laws of physics. And after he did that, many mathematicians and physicians with better math chops pushed his theory and ideas and revealed incredibly rich models and insights and predictions.

How do we get more streams of evidence? By making productive mistakes. By attempting to leverage weird analogies and connections, and iterating on them. We should obviously recognize that most of this will be garbage, but you'll be surprised how many brilliant ideas in the history of science first looked like, or were, garbage.

So if you're worried about AI risk, and want to know if there's anything that can be done, the answer is a resounding yes. There are so many ways of improving our understanding and thus our chances: participating in current research programs and agendas, coming up with new weird takes and approaches, exploring the mechanism, history, and philosophy of science to accelerate the process as much as we can...^[2]

I don't know if we'll make it in time. 5 to 15 years^[3] is a tight deadline indeed, and the strong alignment problem is incredibly complex and daunting. But I know this: if we solve the problem and get out of this alive, this will not be by waiting for an obviously convincing approach; it will come instead from making as many productive mistakes as we can, and learning from them as fast as we can.

1. [△]

I'm not discussing applied alignment research here, like the work of Redwood, but I also find this part crucial and productive. It's just that such work is less about "formulating a solution" and more about "exploring the models and the problems experimentally", which fit well with the model I'm drawing here.

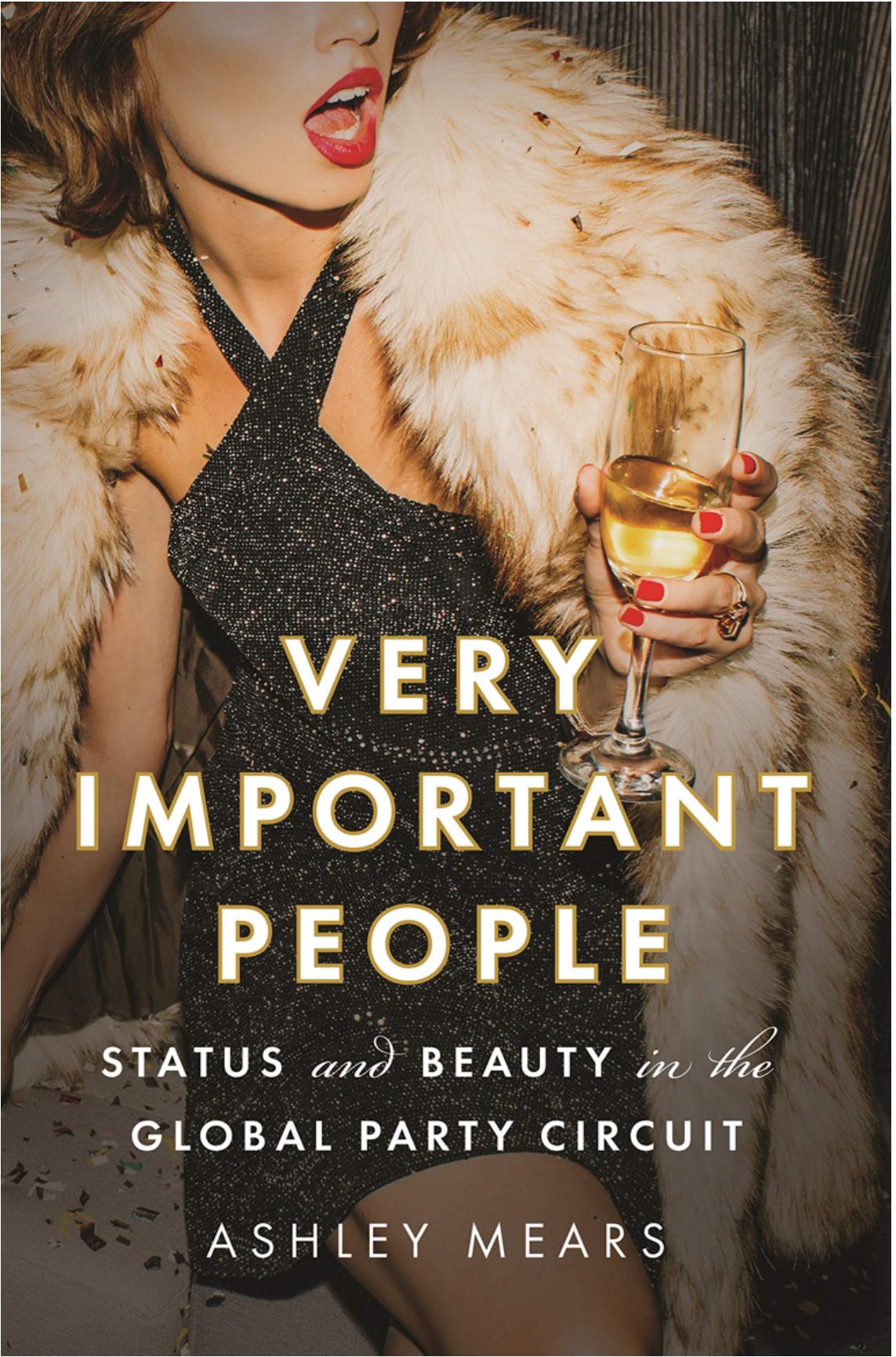
2. [△]

I'm currently finishing a sequence arguing for more pluralism in alignment and providing an abstraction of the alignment problem that I find particularly good for generating new approaches and understanding how all the different takes and perspectives relate.

3. [△]

The range where many short timelines put the bulk of their probability mass.

Book review: Very Important People

A photograph of a woman's face and upper body. She has dark hair, red lipstick, and is wearing a black sequined dress with a large, voluminous white fur collar. She is holding a clear glass filled with a golden liquid, likely champagne, in her right hand. Her nails are painted red, and she is wearing a small ring on her middle finger. The background is dark and textured.

VERY IMPORTANT PEOPLE

STATUS and BEAUTY in the
GLOBAL PARTY CIRCUIT

ASHLEY MEARS

New York's nightclubs are the particle accelerators of sociology: reliably creating the precise conditions under which exotic extremes of status-seeking behaviour can be observed. Ashley Mears documents it all in her excellent book *Very Important People: Status and Beauty in the Global Party Circuit*. A model turned sociology professor, while researching the book she spent hundreds of nights in New York's most exclusive nightclubs, as well as similar parties across the world. The book abounds with fascinating details; in this post I summarise it and highlight a few aspects which I found most interesting.

Here's the core dynamic. There are some activities which are often fun: dancing, drinking, socialising. But they become much more fun when they're associated with feelings of high status. So wealthy men want to use their money to buy the feeling of having high-status fun, by doing those activities while associated with (and ideally while popular amongst) other high-status people, particularly beautiful women.

Unfortunately, explicit transactions between different forms of cultural capital are low-status - it demonstrates that you can't get the other forms directly. So the wealthy men can't just pay the beautiful women to come party with them. Instead an ecosystem develops which sells sufficient strategic ambiguity to allow (self- and other-) deception about the transaction which is taking place, via incorporating a series of middlemen.

Specifically, wealthy men pay thousands at these nightclubs for table charges and "bottle service" - already-expensive alcohol marked up by 5x or much more. The nightclubs pay "promoters" to scout out and bring along dozens of beautiful women each night. Those women get access to an exclusive venue with many wealthy men - but by itself that's not enough to motivate regular attendance, at least not from the prettiest. And most are careful not to ruin their reputations by actually accepting payments from the promoters. Instead, in order to bring enough girls, promoters each need to do a bunch of emotional labour, flirting, relationship-building, and many non-cash payments (food, transport, even accommodation). I'm strongly reminded of Michael Sandel's book [*What Money Can't Buy*](#) - the intuitions about the corrosive effects of money are the same, they're just applied to a much less high-minded setting.

Some interesting features of this system:

- At a top club, a promoter might get paid \$1000 a night to bring out a dozen models or women who look like models. Notably, model-like beauty is much more highly-prized than conventional beauty - e.g. the clubs don't allow access to women who aren't unusually tall. Everyone selects for models even when they don't personally find the model look as attractive, because the fashion industry has established this as the Schelling look for high-status women. (For more on how this happens, see Mears' other book, *Pricing Beauty*; and the responses to [my tweet about it](#)).
- The markup on increasingly large champagne bottles is determined less by the amount of champagne, and more by how ostentatious the purchase is. The biggest purchases, costing over 100k per bottle, therefore come with incredibly elaborate fanfare: all music stops, spotlights shine on the buyer, a whole train of staff bring out the drinks, etc.
- The nightclub profits by creating an atmosphere of "suspended reality" where a large group of people who all individually believe that buying status in this way is tacky can still convince themselves that all the other people *don't* think it's

tacky. Most of the profits don't actually come from the biggest spenders, but rather the next tier down, who are inspired by the atmosphere, and anchored by stories of the biggest purchases.

- In contrast to the predominantly-white clients and models, promoters are disproportionately black. Mears talks about them having "colour capital", and using some stereotypes to their advantage in order to catch attention. They need to be very charismatic and attractive in order to consistently convince girls to come along with them while not making their relationship seem too transactional.
- In some sense the whole system is grounded in the models' sex appeal, but I think that the models' prestige is just as important - as mentioned above, models are preferred to women who most men find more attractive, as well as preferred to women who have more transactional attitudes towards sex.
- Basically the same dynamics play out internationally as well - promoters offer girls free flights, food and accommodation in exchange for attendance at nightclubs in St Tropez, etc. On those trips the transactionality is usually a bit more obvious.
- How can promoters afford to regularly wine and dine so many girls? Often they have deals with restaurants who give them leftover food in exchange for making the restaurant look more glamorous. Other times, wealthy men will host the dinners before the parties start. At the nightclub itself, they all drink for free.

If I were a bit more cynical I might also say that the "fun" part of high-status fun is also mainly a strategic ambiguity which helps facilitate the status transaction - if people couldn't convince themselves and others that they were having fun, their attempts to seem prestigious would be much more obvious. Perhaps it's worth considering what differences you'd expect in a world where this is true vs false. (For example, might you expect that the highest-status men actually don't spend much time dancing, drinking, or even socialising?)

The same might be true, to a lesser extent, of other types of high-status fun - which, in my circles, often involves quick-witted exchanges on arbitrary topics. Overall, though, after reading this book I do feel much luckier that silicon valley is largely disdainful of conspicuous consumption and other negative-sum status games; long may it stay that way.

Greyed Out Options

Imagine that life is a choose-your-own adventure game.

In any moment, you have literally millions of options. At the moment I'm typing this, I could change the tab to innumerable websites, I could read any of the hundreds of books in my house, I could make myself a snack of olives, I could stand up and see how far I could jump, I could pet a cat, I could walk into my best friend's bedroom and call them an idiot, and so on and so forth.

But, most of the time, we only think of a menu of a few dozen options—sometimes much fewer. The rest are sort of grayed out.

To a large extent, this is a good thing. Most of the options theoretically available at any given moment are very stupid. (Just ask anyone with intrusive thoughts—yes, brain, I understand I could put the lightbulb in my mouth, stop bringing it up!) But I think it's important to think about the ways that grayed out options limit our behavior.

You can go outside in pajamas. It isn't illegal. No one will stop you. Most of the time, no one will even comment. Sure, you might run into someone you know, but in many cities that's not going to happen, and anyway they're likely to assume you have a stomach flu or otherwise have some perfectly good reason for running around in pajamas. You're unlikely to face any negative consequences whatsoever.

But when I've suggested this to people, they tend to object not because they have no particular reason to go places in pajamas (pajamas are very comfortable) but because people *don't do that*. It's just not on the list of available options. If you did, you'd probably feel anxious and maybe even ashamed, because it's genuinely hard to do something that people don't do.

To be clear, I'm not suggesting that you should go places wearing pajamas! I don't. I'm suggesting that you consider thoughtfully which of your options are grayed out and why.

Here are some other grayed-out options I've observed among people I've met:

- Starting a conversation with a stranger.
- Asking someone out.
- Eating at a restaurant alone.
- Walking alone at night (especially if you're female or were raised female).
- Writing a novel or a blog post.
- Drawing a picture.
- Submitting your writing to a publisher.
- Emailing a professor to ask a question about their discipline.
- Making a pull request on Github.
- Editing Wikipedia.
- Writing a computer program to fix a problem you have or automate a piece of work you have to do often.
- Starting a recurring event like a bookclub or a meetup.
- Throwing a party.
- Complaining to customer service.
- Opening up a broken machine and poking around in there to see if something obvious is wrong.

- Googling your problem.
- Negotiating your salary.
- Researching a topic of interest on Google Scholar.
- Doing parkour on walls etc that you find on your walk.
- Planting potato eyes etc and getting food from them.
- Talking to a famous person.
- Singing in public.
- Transitioning.
- Dating people of the same gender.
- Talking openly with your partner about your relationship needs.
- Traveling the world and crashing on various friends' couches instead of having a house.
- Cutting off your family.
- Being out about something stigmatized (your disability, your sexual orientation, your religion, your hobbies...).
- Asking for your unusual preferences to be accommodated (by service workers or by people you know).

Different people have different grayed-out options, and I think this is actually a *really* common reason that people behave differently from each other. The reason that I write blog posts and other people don't is not that I'm good at writing and they're not; it's that writing a blog post about something I'm thinking about is on my action menu and it's not on theirs.

Some grayed-out options just don't occur to you (like eating a lightbulb, if you are more blessed than I am). But I think there are two distinct feelings associated with an option that does occur to you still being grayed out.

First, there's a feeling of *social judgment*. Even if you're never going to see those people again and all of them are too wrapped up in their own business to pay attention to you anyway, if you're pursuing a grayed-out option, it's very common to feel like there are dozens of eyes watching you. It feels like everyone is whispering behind your back saying "who does that?" You can even feel this sense internalized when you're doing something by yourself with no one watching.

Second, there's a feeling of *overwhelm*. This is particularly common with the more complex grayed-out options: the steps involved in eating alone at a restaurant are obvious to nearly everyone. But thinking about a grayed-out option like editing Wikipedia, asking someone out, doing a pull request, or writing a blog post can be very overwhelming.

For me, writing a blog post is a series of concrete steps: notice thoughts I'm having, talk my idea over with a friend to get a sanity check, open my Substack word processor, type up an explanation of my thoughts, reread it to make sure it makes sense, send it to my betas for a sanity check, do a last proofread for typos and grammatical mistakes, and schedule it for Tuesday or Friday morning. If writing a blog post is a grayed-out action for you, it's more like an amorphous mass of question marks. Even if you tried, you'd probably just stare at a blinking cursor with no idea what to do next. *What are you even supposed to be doing?*

Crucially, something can be a non-grayed-out option even if you've never done it. At one point I'd never organized a LARP, but it was clearly an option long before I did it. "Organize a LARP" was an action made of a series of steps, like "pick a LARP to run" and "find friends who would want to be in a LARP or who are susceptible to my puppy-

“dog eyes” and “find a venue” and “nag people about how they really do have to finish their character sheets.”

I think options are less likely to be grayed out if they’re similar to something you know how to do. If you’ve organized themed parties before, organizing a LARP is more likely to feel like an option; if you’ve written fiction, writing blog posts is more likely to feel like an option; if you’ve worn cosplay to the store, wearing pajamas is more likely to feel like an option. For more complex options in particular, similarity to things you’ve done before helps, because the overall option is made up of a bunch of smaller things that are *individually options*. Running a LARP is easier if you already have “find a venue” and “make puppy-dog eyes at friends” on your action menu.

They’re also less likely to be grayed out if you know people who do it. That’s the truth behind “queerness is a social contagion”: if people you know are transitioning genders or dating people of the same gender, suddenly that puts those options on your action menu too. To some extent, knowing people with an option on their action menu helps because you have someone to direct your stupid questions to and receive advice from. But I think a surprising amount is just the fact that *other people are clearly doing it*. This is a thing real humans you really know do, not a thing theoretically done by people far away in Hollywood or Washington D. C. or Timbuktu.

Again, I’m not some kind of action-menu maximalist. I don’t think anyone has every possible action on their action menu, even if you just look at things people might want to do like organizing LARPs and not things that no one wants to do like eating lightbulbs. But I think that it’s pretty common that there are things people could do that they want to do or that would really improve their life if they did, and nothing is *actually stopping them*. It just... doesn’t feel to them like the sort of thing that people do. And I think it’s useful to look at your action menu and see if there are some grayed-out options there you want to make colorful.

Giving calibrated time estimates can have social costs

In the normal course of my rationalist upbringing, I learned about the classic cognitive biases, including the [planning fallacy](#). This is essentially the fact that it almost always takes longer for a task to get done than people estimate at the beginning. The explanation described in the original [sequence post](#) is that people visualize the mainline path of accomplishing the steps of the task, and then just add those times together, whereas in reality, at least one of those steps will have something go wrong, and make the whole thing take much longer.

So I read this, and then just... updated? Typically one should be very skeptical about the feeling that they are not subject to a bias, but as discussed in the original sequence post, this one is empirically reported to be correctable. And I'm sure it took me some time to adjust, but it wasn't too difficult to install a [TAP](#) to just consult the outside view instead of the inside view.

And then I went forth into the world, and gave calibrated time estimates, because I want to have true beliefs and I want to say true things. Slowly, over time, I got the sinking sense that this was costing me social points.

Types of costs

People may think you're slow

A primary context in which I had to give time estimates was when working as a software engineer. I've worked at multiple companies that regularly used "sprints" to plan work, and we often assigned explicit estimates to the tasks at hand. I regularly gave longer estimates than my coworkers or manager. These estimates were also calibrated; I actually took about that long. Sometimes more, sometimes less. My coworkers usually took longer than their estimates, but their estimates were shorter.^[1]

Over time, I believe my managers got a [worse impression](#) of me.^[2] One simple problem here is that shorter estimates just sound better than longer estimates because the manager wants the thing done sooner. But another slightly subtler problem is that since almost everyone gives optimistic estimates, my manager would reasonably be used to hearing optimistic estimates. So if they hear me give a longer estimate, they don't know that it's more calibrated, so they subconsciously assume it's just as optimistic as my coworkers, and then they would reasonably believe that I am worse.

Ideally, my calibration would eventually be evident through cumulative statistics. But realistically, people aren't paying attention on this level. They regularly feel the short-term displeasure of longer estimates, and they only irregularly get the pleasure from noticing that I finished it earlier than I said, and it's hard to connect up these two signals that are days or weeks apart.

People may think you're lazy

Claiming that things will take longer also just sounds like you're trying less hard. If Jane says she can get this done in a day, why can't you? Are you saying she's lying? Maybe you're just not trying as hard.

People dislike pessimism

Saying that something will take longer is negative-valence. You are claiming that the world is worse, and hearing that feels bad. People don't like regulations and safety expectations, because in the short term they are annoying and costly, even though in the long term they are designed to be much better for you. Hearing someone say that something will take longer is annoying and costly in the short term, even though being calibrated could allow the whole project to plan better.

People can feel insulted

The planning fallacy also applies to other people telling me their time estimates. I have no reason to assume other people are calibrated, and in my experience they usually do take longer than they estimate. And yet, telling them that you think their estimate is too short is rarely going to be taken well. Often I can just hear their estimate, not say anything, and then personally assume and prepare for it to take much longer. But that's not always an option, for example if you're working together on the task, or if you care about helping the person plan better

People want commitment

In a related problem, sometimes people will ask me "can you get this done by [time]?" and I'll say something like "seems reasonable" but as the discussion carries on, it becomes clear that they want me to basically *promise* that I'll get the task done by then. My problem with this has two parts. One is that my priors on how long it takes to do things in general are not just later than others, but also very wide. There just is a good chance that things will take *much* longer.

The second part is that, for me to utter a statement X, I have to believe it with a certain probability that is fairly high. So when someone wants me to echo back the unqualified statement "I will get it done by [time]", they're basically asking me to be, I dunno, 98% confident, and I just can't be, because I know that the 98th percentile time is actually more like $3 \times [\text{time}]$. So I continue to hedge, and then if they push I basically tell them the above, and then they feel like I'm making a point to not commit, and then maybe they start thinking the above things, like that I'm slow or lazy or something.

The annoying thing here is that I believe the only difference between me and another task doer in this situation is that I have more accurate beliefs, or I have a higher belief threshold for making claims (or something similar, like that I only use statements for communicating beliefs and not for socially enforcing a commitment to myself). I think I can get the thing done just as fast as someone else, and thus I think I can satisfy the task-giver just as much, but they currently feel less satisfied because of the above communication.

Being out of sync with others

The last type of cost I can think of is when me and someone else are both equal parties subject to something else getting done. Say that me, Jane and Jack are going to the movies. I know that Jack always takes a long time to get ready at the last minute, and so I am psychologically prepared for this. We show up at Jack's house, and I go ahead and sit down on the couch and start reading on my phone. Maybe we'll miss the first few minutes of the movie, or get worse seats or something, but I've already accepted that as a likely consequence of going to the movies with Jack. Jane however has not done the same level of planning fallacy compensation as me, and so Jane feels surprised and frustrated when Jack takes a long time. This isn't exactly a conflict between me and Jane, but it can be awkward. Maybe Jane is anxiously pacing around while I'm relaxing on the couch. Maybe she feels like I should be on her side by feeling equally impatient, or something. Maybe she feels like I'm enabling Jack by not demonstrating urgency.

An example that did happen to me is that my house is having its electrical system replaced. This is taking a very, very, very long time. Many of my roommates seem not just annoyed (which anyone reasonably would be) but also something like surprised by this. And, I dunno, I just knew it would take forever and ever? I just decided to start living my life as if it would never end. I remember literally thinking to myself, "it would be nice if this got finished in 2021". I currently have my fingers crossed for it being finished in 2022.

And again, this isn't exactly a conflict between me and my roommates, but I feel a little bit snobbish just for even writing the above sentences. And I think that difference can seed some kind of low-key tension after enough repetitions.

Solutions?

Mostly my solution to the above is to just remember that they happen, and add that in to the social calculus that I'm always running every time I interact with a person. Often I just pay the social costs, partly because I have pretty high social resilience, partly because I continue to want to be a person that acts from my beliefs even when it's costly. (The solution of "just give shorter time estimates" is basically a non-starter for me.) The adjustments I do make are usually in the form of giving the estimate with softer wording, some kind of hedging, or just being more clear about my intention versus my uncertainty. I'm pretty sure that a more socially skilled person than me would have clearly communicable solutions that used some kind of charisma warmth to make the listener feel better despite the longer estimate.

1. ^

Here and elsewhere, feel free to just, like, not believe my self-reports. I do not in fact have the data handy in spreadsheets, or anything.

2. ^

To be totally fair, there were also times when I was actually worse than my coworkers. But this problem compounded the impression during those times.

Code Generation as an AI risk setting

Historically, it has been difficult to persuade people of the likelihood of AI risk because the examples tend to sound “far-fetched” to audiences not bought in on the premise. One particular problem with many traditional framings for AI takeover is that most people struggle to imagine how e.g. “a robot programmed to bake maximum pies” figures out how to code, locates its own source-code, copies itself elsewhere via an internet connection and then ends the world.

There’s a major logical leap there: “pie-baking” and “coding” are things done by different categories of agent in our society, and so it’s fundamentally odd for people to imagine an agent capable of both. This oddness makes it feel like we must be far away from any system that could be *that* general, and thus pushes safety concerns to a philosophical exercise.

I want to make the case that the motivating example we should really be using is automatic code generation. Here’s a long list of reasons why:

- It’s obvious to people why and how a system good at generating code could generate code to copy itself, if it were given an open-ended task. It’s a basic system-reliability precaution that human engineers would also take.
- For non-experts, they are already afraid of unrestrained hackers and of large tech companies building software products that damage society - this being done by an unaccountable AI fits into an emotional narrative.
- For software people (whom we most need to convince) the problem of unexpected behaviors from code is extremely intuitive - as is the fact that it is always the case that code bases are too complex for any human to be certain of what they’ll do before they’re run.
- Code generation does seem to be getting dramatically better, and the memetic/media environment is ripe for people to decide how to feel about these capabilities.
- Nearly all conceivable scalable prosaic alignment solutions will require some degree of “program verification” - making sure that code isn’t being run with an accidentally terrible utility function, or to verify the outputs of other AIs via code-checking Tool AIs. So we want substantial overlap between the AI safety and AI codegen communities.
- The “alignment problem” already exists in nearly all large software engineering projects: it’s very difficult to specify what you want a program to do ahead of time, and so we mostly just run codebases and see what happens.
- All of the concerns around “the AI learns to use Rowhammer to escape” feel much more obvious when you’re building a code-generator.
- We can even motivate the problem by having the AI’s objective be “make sure that other code-generating AIs don’t misbehave”. This is open-ended in a way that obviously makes it a utility-maximizer, and preemptively addresses the usual technooptimistic response of “we’ll just build auditor AIs” by starting with aligning those as the premise.
- The distinction between act-based AIs and EUMs is obvious in the case of code-gen. Similarly, the idea of Safety via Debate is related to code reviewing processes.
- Software project generation capabilities seem both necessary and possibly sufficient for FOOM/takeover scenarios.

- Ultimately, the people in government/companies most sympathetic to high-tech risk mitigation are the people who think about cybersecurity - so scaring them gets us a very useful ally. (It's also a community with plenty of people with the "security mindset" needed for many empirical alignment scenarios.)

On the other hand, there may be some risk that focusing on code generation increases its public salience and thus investment in it. But this seems likely to have happened anyway. It's also more obviously the path towards recursive self-improvement, and thus may accelerate AI capabilities, but again this does already seem to be happening whether or not we discuss it.

What do people think of this as a framing device?

[Closed] Hiring a mathematician to work on the learning-theoretic AI alignment agenda

UPDATE: The position is now closed. My thanks to everyone who applied, and also to those who spread the word.

The Association for Long Term Existence and Resilience (ALTER) is a new charity for promoting longtermist^[1] causes based in Israel. The director is David Manheim, and I am a member of the board. Thanks to a generous grant by the [FTX Future Fund](#) Regrants Program, we are recruiting a researcher to join me in working on the [learning-theoretic research agenda](#)^[2]. The position is remote and suitable for candidates in most locations around the world.

Apply [here](#).

Requirements

- The candidate must have a track record in mathematical research, including proving non-trivial original theorems.
- The typical candidate has a PhD in theoretical computer science, mathematics, or theoretical physics. However, we do not require the diploma. We do require the relevant knowledge and skills.
- Background in one or several of the following fields is an advantage: statistical/computational learning theory, algorithmic information theory, computational complexity theory, functional analysis.

Job Description

The researcher is expected to make progress on open problems in the learning-theoretic agenda. They will have the freedom to choose any of those problems to work on, or come up with their own research direction, as long as I deem the latter sufficiently important in terms of the agenda's overarching goals. They are expected to achieve results with minimal or no guidance. They are also expected to write their results for publication in academic venues (and/or informal venues such as the alignment forum), prepare technical presentations et cetera. (That said, we rate researchers according to the estimated impact of their output on reducing AI risk, not according to standard academic publication metrics.)

Here are some open problems from the agenda, described very briefly:

- Study the mathematical properties of the algorithmic information-theoretic [definition of intelligence](#). Build and analyze formal models of value learning based on this concept.
- Pursue any of the [future research directions](#) listed in the article on infra-Bayesian physicalism.
- Continue the study of reinforcement learning with [imperceptible rewards](#).

- Develop a theory of quantilization in reinforcement learning (building on the corresponding [control theory](#)).
- Study the [overlap](#) of algorithmic information theory and statistical learning theory.
- Study [infra-Bayesian logic](#) in general, and its applications to infra-Bayesian reinforcement learning in particular.
- Study the behavior of RL agents in [population games](#). In particular, understand to what extent [infra-Bayesianism](#) helps to avoid the [grain-of-truth problem](#).
- Develop a theory of [antitraining](#): preventing AI systems from learning particular domains while learning other domains.
- Study the [infra-Bayesian Turing reinforcement learning](#) setting. This framework has applications to reflective reasoning and hierarchical modeling, among other things.
- Develop a theory of reinforcement learning with [traps](#), i.e. irreversible state transitions. Possible research directions include studying the computational complexity of Bayes-optimality for finite state policies (in order to avoid the [NP-hardness](#) for arbitrary policies) and [bootstrapping](#) from a safe baseline policy.

Terms

The position is full-time, and the candidate must be available to start working in 2022. The salary is between 60,000 USD/year to 180,000 USD/year, depending on the candidate's prior track record. The work can be done from any location. Further details depend on the candidate's country of residence.

1. Personally, I don't think the long-term future should override every other concern. And, I don't consider existential risk from AI especially "long term" since it can plausibly materialize in my own lifetime. Hence, "longtermist" is better understood as "important even if you *only* care about the long-term future" rather than "important *only* if you care about the long-term future". [←](#)
2. The linked article is not very up-to-date in terms of the open problem, but is still a good description on the overall philosophy and toolset. [←](#)

Clem's Memo

Declassified document sourced from Cabinets and The Bomb by Peter Hennessy; reproduced verbatim.

TOP SECRET

GEN 75/1

28th August, 1945

THE ATOMIC BOMB

Memorandum by the Prime Minister [\[1\]](#)

1. A decision on major policy with regard to the atomic bomb is imperative. Until this is taken civil and military departments are unable to plan. It must be recognised that the emergence of this weapon has rendered much of our post-war planning out of date.
2. For instance a redistribution of industry planned on account of the experience of bombing attacks during the war is quite futile in the face of the atomic bomb. Nothing can alter the fact that the geographical situation of Britain offers to a Continental Power such targets as London and the other great cities. Dispersal of munition works and airfields cannot alter the facts of geography.
3. Again it would appear that the provision of bomb proof basements in factories and offices and the retention of A.R.P.[\[2\]](#) and Fire Services is just futile waste.
4. All considerations of strategic bases in the Mediterranean or the East Indies are obsolete. The vulnerability of the heart of the Empire is the one fact that matters. Unless its safety can be secured, it is no use bothering about things on the periphery. It is difficult for people to adjust their minds to an entirely new situation. I noticed at Potsdam that people still talked of the line of the Western Neisse although rivers as strategic frontiers have been obsolete since the advent of Air Power. It is infinitely harder for people to realise that even the modern conception of war to which in my lifetime we have become accustomed is now completely out of date.
5. We recognise or some of us did before this war that bombing could only be answered by counter bombing. We were right. Berlin and Magdeburg were the answer to London and Coventry. Both derive from Guernica. The answer to an atomic bomb on London is an atomic bomb on another great city.
6. Duelling with swords and inefficient pistols was bearable. Duelling had to go with the advent of weapons of precision. What is to be done about the atomic bomb? It has been suggested that by a Geneva Convention all nations might agree to abstain from its use. This method is bound to fail as it has failed in the past. Gas was forbidden but used in the first world war. It was not used in World War 2, but its belligerents were armed with it. We should have used it, if the Germans had landed on our beaches. It was not used, because military opinion considered it less effective than explosives and incendiaries.
7. Further the banning of the atomic bomb would leave us with the other weapons used in the late war which were quite destructive enough.
8. Scientists agree that we cannot stop the march of discovery. We can assume that any attempt to keep this as a secret in the hands of the U.S.A. and U.K. is useless. Scientists in other countries are certain in time to hit upon the secret.

9. The most we may have is a few years start. The question is what use we are to make of that few years start.
10. We might presumably on the strength of our knowledge and of the advanced stage reached in technical development in the U.S.A. seek to set up an Anglo-American Hegemony in the world using our power to enforce a world wide rigid inspection of all laboratories and plants.
11. I do not think this is desirable or practicable. We should not be able to penetrate the curtain that conceals the vast area of Russia. To attempt this would be to invite a world war leading to the destruction of civilization in a dozen years or so.
12. The only course which seems to me to be feasible and to offer a reasonable hope of staving off disaster for the world is joint action taken by the U.S.A., U.K. and Russia based on stark reality.
13. We should declare that this invention has made it essential to end wars. The new World Order must start now. The work of the San Francisco Conference must be carried much further.
14. While steps must be taken to prevent the development of this weapon in any country, this will be futile unless the whole conception of war is banished from people's minds and from the calculations of governments. This means that every vexed question will have to be settled without the use of force, whether it is Palestine, Venezia Giulia, the Ruhr, India. Every nation must submit to the rule of law. The U.S.S.R. must abandon, if it still holds them, its dreams of revolution by force or intrigue. The U.K. and the U.S.A. must abandon, if they have them, any dreams of overturning Left Governments. All nations must give up their dreams of realizing some historic expansion at the expense of their neighbours. They must look to a peaceful future instead of a warlike past.
15. This sort of thing has been considered a Utopian dream. It has become today the essential condition of the survival of civilisation and possibly of life in [3] this planet.
16. No government has ever been placed in such a position as is ours today. The Governments of the U.K. and the U.S.A. are responsible as never before for the future of the human race.
17. I can see no other course than that I should on behalf of the Government put the whole of the case to President Truman and propose that he and I and Stalin should forthwith take counsel together.

The time is short.

We must come to a decision before the meeting of the United Nations Organization.

We cannot plan our future while this major factor is uncertain.

I believe that only a bold course can save civilization.

1. ^

UK Prime Minister Clement Attlee, shortly after succeeding Winston Churchill.

2. ^

Air Raid Precautions

3. ^

[sic]

Intuitions about solving hard problems

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

Solving hard scientific problems usually requires compelling insights

Here's a heuristic which plays an important role in my reasoning about solving hard scientific problems: that when you've made an important breakthrough, you should be able to explain the key insight(s) behind that breakthrough in an intuitively compelling way. By "intuitively compelling" I don't mean "listeners should be easily persuaded that the idea solves the problem", but instead: "listeners should be easily persuaded that this is the type of idea which, if true, would constitute a big insight".

The best examples are probably from Einstein: time being relative, and gravity being equivalent to acceleration, are both insights in this category. The same for Malthus and Darwin and Gödel; the same for Galileo and Newton and Shannon.

Another angle on this heuristic comes from Scott Aaronson's [list of signs that a claimed \$P \neq NP\$ proof is wrong](#). In particular, see:

#6: the paper lacks a coherent overview, clearly explaining how and why it overcomes the barriers that foiled previous attempts.

And #1: the author can't immediately explain why the proof fails for 2SAT, XOR-SAT, or other slight variants of NP-complete problems that are known to be in P.

I read these as Aaronson claiming that a successful solution to this very hard problem is likely to contain big insights that can be clearly explained.

Perhaps the best counterexample is the invention of Turing machines. Even after Turing explained the whole construction, it seems reasonable to still be uncertain whether there's actually something interesting there, or whether he's just presented you with a complicated mess. I think that uncertainty would be particularly reasonable if we imagine trying to understand the formalism *before* Turing figures out how to implement any nontrivial algorithm (like prime factorisation) on a Turing machine, or how to prove any theorems about universal Turing machines.

Other counterexamples might include quantum mechanics, where quantization was originally seen as a hack to make the equations work; or formal logic, where I'm not sure if there were any big insights that could be grasped in advance of actually seeing the formalisms in action.

Using the compelling insight heuristic to evaluate alignment research directions

It's possible that alignment will in practice end up being more of an *engineering* problem than a scientific problem like the ones I described above. E.g. perhaps we're in a world where, with sufficient caution about scaling up existing algorithms, we'll produce aligned AIs capable of solving the full version of the problem for us.^[1] But suppose we're trying to produce a fully scalable solution ourselves; are there existing

insights which might be sufficient for that? Here are some candidates, which I'll only discuss very briefly, and plan to discuss in more detail in a forthcoming post (I'd also welcome suggestions for any I've missed):

- “Trustworthy imitation of human external behavior would avert many default dooms as they manifest in external behavior unlike human behavior.”
 - This is [Eliezer's description](#) of the core insight behind Paul's imitative amplification proposal. I find this somewhat compelling, but less so than I used to, since I've realized that the line between imitation learning and reinforcement learning is blurrier than I used to think (e.g. see [this](#) or [this](#)).
- Decomposing supervision of complex tasks allows better human oversight.
 - Again, I've found this less compelling over time - in this case because I've realized that decomposition is the “default” approach we follow whenever we evaluate things, and so the real “work” of the insight needs to be in describing *how* we'll decompose tasks, which I don't think we've made much progress on (with techniques like cross-examination being possible exceptions).
- Weight-sharing makes deception much harder.
 - I think this is the main argument pushing me towards optimism about ELK; thanks to Ajeya for articulating it to me.
- Uncertainty about human preferences makes agents corrigible.
 - This is Stuart Russell's claim about why assistance games are a good potential solution to alignment; I basically don't buy it at all, [for the same reasons as Yudkowsky](#) (but kudos to Stuart for stating the proposed insight clearly enough that the disagreement is obvious).
- Myopic agents can be capable while lacking incentives for long-term misbehavior.
 - This claim seems to drive a bunch of Evan Hubinger's work, but I don't buy it. In order for an agent's behavior to be competent over long time horizons, it needs to be doing some kind of cognition aimed towards long time horizons, and we don't know how to stop that cognition from being goal-directed.
- Problems that arise in limited-data regimes (e.g. inner misalignment) go away when you have methods of procedurally generating realistic data (e.g. separable world-models).
 - This claim was made to me by Michael Cohen. It's interesting, but I don't think it solves the core alignment problem, because we don't understand cognition well enough to efficiently factor out world-models from policies. E.g. training a world-model to predict observations step-by-step seems like it loses out on all the benefits of thinking in terms of abstractions; whereas training it just on long-term predictive accuracy makes the intermediate computations uninterpretable and therefore unusable.
- By default we'll train models to perform bounded tasks of bounded scope, and then achieve more complex tasks by combining them.
 - This seems like the core claim motivating Eric Drexler's CAIS framework. I think it dramatically underrates the importance of general intelligence, and the returns to scaling up single models, for [reasons I explain further here](#).
- Functional decision theory.
 - I don't think this directly addresses the alignment problem, but it feels like the level of insight I'm looking for, in a related domain.

Note that I do think each of these claims gestures towards interesting research possibilities which might move the needle in worlds where the alignment problem is easy. But I don't think any of them are sufficiently powerful insights to scalably solve

the hard version of the alignment problem. Why do many of the smart people listed above think otherwise? I think it's because they're not accounting properly for the sense in which the alignment problem is an adversarial one: that by default, optimization which pushes towards general intelligence will also push towards misalignment, and we'll need to do something unusual to be confident we're separating them. In other words, the set of insights about consequentialism and optimization which made us worry about the alignment problem in the first place (along with closely-related insights like the orthogonality thesis and instrumental convergence) are sufficiently high-level, and sufficiently robust, that unless you're guided by other powerful insights you're less likely to find exceptions to those principles, and more likely to find proposals where you can no longer spot the flaw.

This claim is very counterintuitive from a ML perspective, where loosely-directed exploration of new algorithms often leads to measurable improvements. I don't know how to persuade people that applying this approach to alignment leads to proposals which are deceptively appealing, except by getting them to analyze each of the proposals above until they convince themselves that the purported insights are insufficient to solve the problem. Unfortunately, this is very time-consuming. To save effort, I'd like to promote a norm for proposals for alignment techniques to be very explicit about *where the hard work is done*, i.e. which part is surprising or insightful or novel enough to make us think that it could solve alignment even in worlds where that's quite difficult. Or, alternatively, if the proposal is only aimed at worlds where the problem is relatively easy, please tell me that explicitly. E.g. I spent quite a while being confused about which part of the ELK agenda was meant to do the hard work of solving ontology identification; after asking Paul about it, though, his main response was "maybe ontology identification is easy", which I feel very skeptical about.^[2] (He also suggested something to do with the structure of explanations as a potential solving-ELK-level-insight; but I don't understand this well enough to discuss it in detail.)

Using the compelling insight heuristic to generate alignment research directions

If we need more major insights about intelligence/consequentialism/goals to solve alignment, how might we get them? Getting more evidence from seeing more advanced systems will make this much easier, so one strategy is just to keep pushing on empirical and engineering work, while keeping an eye out for novel phenomena which might point us in the right directions.^[3]

But for those who want to try to generate those insights directly, some tentative options:

- Studying how existing large models think, and try to extrapolate from that
- Understand human minds well enough to identify insights about human intelligence (e.g. things like predictive coding, or multi-agent models of minds, or dual process theory) which can be applied to alignment
- Understanding how groups think (e.g. how task decomposition occurs in cultural evolution, or in corporations, or...)
- Agent foundations research

Of course, the whole point of major insights is that it's hard to predict them; so I'd be excited about others pursuing potential insights that don't fall into any of these categories (with the caveat that, as the work gets increasingly abstract, it's necessary to be increasingly careful for it to have any chance of succeeding).

1. ^

I model Eliezer as agreeing with most of the claims I make in this post, but strongly disagreeing with this sentence, because he thinks that the core problem is so hard that no amount of prosaic engineering effort could plausibly prevent catastrophe in the absence of major novel insights.

2. ^

Some brief intuitions about why: I think the hardest part of human cognition is generating and merging different ontologies. Thinking “within” an ontology is like doing normal research in a scientific field; reasoning about different ontologies is like doing philosophy, or doing paradigm-breaking research, and so it seems like a particularly difficult thing to generate a training signal for.

3. ^

Thanks to Nathan Helm-Burger for reminding me of this, with his comment.

Good Heart Week: Extending the Experiment

The Good Heart Project

♥ Goodest Hearts	♥ Good Hearts	♥ Kinda Good Hearts
1 lsusr \$246	6 johnswentworth \$119	11 gwern \$82
2 Eliezer Yudkowsky \$201	7 NunoSempere \$119	12 Matthew Barnett \$77
3 Ben Pace \$180	8 hath \$113	13 Linch \$58
4 Vaniver \$129	9 abramdemski \$89	14 jefftk \$57
5 G Gordon Worley III \$121	10 AprilSR \$85	15 jessicata \$54

[What's this about?](#)

The Good Heart Leaderboard, at the time of writing.

Yesterday we launched [Good Heart Tokens](#), and said they could be exchanged for 1 USD each.

Today I'm here to tell you: ***this is actually happening and it will last a week. You will get a payout if you give us a PayPal/ETH address or name a charity of your choosing.***

Note that voting rings and fundraising are now out of scope, we will be removing and banning users who do that kind of thing starting now. More on this at the end of the post.

Also, we're tentatively changing posts to be worth 4x the Good Heart Tokens of comments (*Update: we decided on 3x instead, and it just went live, at around 4:20PM PT on April 2nd*).

Why is this experiment continuing?

Let me state the obvious: if this new system were to last for many months or years, I expect these financial rewards would change the site culture for the worse. It would select on pretty different motives for being here, and importantly select on different people who are doing the voting, and then the game would be up.

(Also I would spend a lot of my life catching people explicitly trying to game the system.)

However, while granting this, I suspect that in the *short run* giving LessWrong members and lurkers a stronger incentive than usual to write well-received stuff has the potential to be great for the site.

For instance, I think the effect yesterday on site regulars was pretty good. I'll quote AprilSR [who said:](#)

I am not very good at directing my monkey brain, so it helped a lot that my System 1 really anticipated getting money from spending time on LessWrong today.

...There's probably better systems than "literally give out \$1/karma" but it's surprisingly effective at motivating me in particular in ways that other things which have been tried very much aren't.

I think lots of people wrote good stuff, much more than a normal day. Personally my favorite thing that happened due to this yesterday was when people published a bunch of their drafts that had been sitting around, some of which I thought were excellent. I hope this will be a kick for many people to actually sit down and write that post they've had in their heads for a while.

(I certainly don't think money will be a motivator for *all* people, but I suspect it is true for *enough* that it will be worth it for us given the Lightcone Infrastructure team's value of money.)

I'm really interested to find out what happens over a week, I have a hope it will be pretty good, and the Lightcone Infrastructure team has the resources that makes the price worth it to us. So I invite you into this experiment with us :)

Info and Rules

Here's the basic info and rules:

- **Date:** Good Heart Tokens will continue to be accrued until EOD Thursday April 7th (Pacific Time). I do not expect to extend it beyond then.
- **Scope:** We are no longer continuing with "fun" uses of the karma system. Voting rings, fundraising posts, etc, are no longer within scope. Things like John Wentworth's and Aphyer's [voting ring](#), and G Gordon Worley III's [Donation Lottery](#) were both playful and fine uses of the system on April 1st, but from now I'd like to ask these to stop.
- **Moderation:** We'll bring mod powers against accounts that are abusing the system. We'll also do a pass over the votes at the end of the week to check for any suspicious behavior (while aiming to minimize any deanonymization).
- **Eligible:** LW mods and employees of the Center for Applied Rationality are not eligible for prizes.
- **Votes:** Reminder that only votes from pre-existing accounts are turned into Good Heart Tokens. (But new accounts can still earn tokens!) And of course self-votes are not counted.
- **Cap Change:** We're lifting the 600 token cap to 1000. (If people start getting to 1000, we will consider raising it further, but no promises.)
- **Weight Change:** We're tentatively changing it so that votes on posts are now worth 4x votes on comments. (*Update: we decided on 3x instead, and it just went live, at around 4:20PM PT on April 2nd.*)
- **Getting Money:** To receive your funds, please log in and enter your payment info at lesswrong.com/payments/account. The minimum amount you can get is \$25.
- **Catch-all:** From some perspectives this is a crazy experiment, so I want to acknowledge up-front that if something pretty bad and unexpected happens we'll do what seems best to us. Nothing is certain, and there are some worlds where we don't end up paying out things that some of you had hoped. We'll adapt as we go.

Go forth and write excellent posts and comments!

Convince me that humanity is as doomed by AGI as Yudkowsky et al., seems to believe

I've been very heavily involved in the (online) rationalist community for a few months now, and like many others, I have found myself quite freaked out by the apparent despair/lack of hope that seems to be sweeping the community. When people who are smarter than you start getting scared, it seems wise to be concerned as well, even if you don't fully understand the danger. Nonetheless, it's important not to get swept up in the crowd. I've been trying to get a grasp on *why* so many seem so hopeless, and these are the assumptions I believe they are making (trivial assumptions included, for completeness; there may be some overlap in this list):

1. AGI is possible to create.
2. AGI will be created within the next century or so, possibly even within the next few years.
3. If AGI is created by people who are not sufficiently educated (aka aware of a solution to the Alignment problem) and cautious, then it will almost certainly be unaligned.
4. Unaligned AGI will try to do something horrible to humans (not out of maliciousness, necessarily, we could just be collateral damage), and will not display sufficiently convergent behavior to have anything resembling our values.
5. We will not be able to effectively stop an unaligned AGI once it is created (due to the Corrigibility problem).
6. We have not yet solved the Alignment problem (of which the Corrigibility problem is merely a subset), and there does not appear to be any likely avenues to success (or at least we should not expect success within the next few decades).
7. Even if we solved the Alignment problem, if a non-aligned AGI arrives on the scene before we can implement ours, we are still doomed (due to first-mover advantage).
8. Our arguments for all of the above are not convincing or compelling enough for most AI researchers to take the threat seriously.
9. As such, unless some drastic action is taken soon, unaligned AGI will be created shortly, and that will be the end of the world as we know it.

First of all, is my list of seemingly necessary assumptions correct?

If so, it seems to me that most of these are far from proven statements of fact, and in fact are all heavily debated. Assumption 8 in particular seems to highlight this, as if a strong enough case could be made for each of the previous assumptions, it would be fairly easy to convince most intelligent researchers, which we don't seem to observe.

A historical example which bears some similarities to the current situation may be Godel's resolution to Hilbert's program. He was able to show unarguably that no consistent finite system of axioms is capable of proving all truths, at which point the mathematical community was able to advance beyond the limitations of early formalism. As far as I am aware, no similarly strong argument exists for even one of the assumptions listed above.

Given all of this, and the fact that there are so many uncertainties here, I don't understand why so many researchers (most prominently Eliezer Yudkowsky, but there are countless more) seem so certain that we are doomed. I find it hard to believe that all alignment ideas presented so far show no promise, considering I've yet to see a slam-dunk argument presented for why even a single modern alignment proposals *can't* work. (Yes, I've seen proofs against straw-man proposals, but not really any undertaken by a current expert in the field). This may very well be due to my own ignorance/ relative newness, however, and if so, please correct me!

I'd like to hear the steelmanned argument for why alignment is hopeless, and Yudkowsky's announcement that "I've tried and couldn't solve it" without more details doesn't really impress me. My suspicion is I'm simply missing out on some crucial context, so consider this thread a chance to share your best arguments for AGI-related pessimism. (Later in the week I'll post a thread from the opposite direction, in order to balance things out).

EDIT: Read the comments section if you have the time; there's some really good discussion there, and I was successfully convinced of a few specifics that I'm not sure how to incorporate into the original text. ☺

Working Out in VR Really Works

Epistemic Status: **Based almost entirely on my own experience.** Also based on some friends [here](#), [here](#), and in real life.

TL;DR: I now need willpower in order to not-workout (!). I wish someone would have told me this a long time ago when I was looking for some workout that would resonate with me.

Is it actually a workout?

TL;DR: Yes.

1. I get tired and I sweat, especially in a specific intensive game (Thrill of the Fight, aka TotF)
2. I saw someone play that game for under 3 minutes and collapse on the sofa in exhaustion for about half a day
3. Here's a graph from my activity tracker (Oura Ring gen 2, **not optimal for tracking workouts!**, but still adds some info and also everyone likes graphs, right?) :



To give context to this graph (of me playing TotF several times during the day): When I do a rollerblade trip with a group that's better than me, the **hardest** part is usually under the "high" line. TotF is often above it.

Can you build muscle with VR?

TL;DR: Unclear.

People who know stuff (I don't qualify) say one needs resistance to build muscles, so maybe using wrist-weights or elastic bands that connect one's limbs to one's torso. I didn't try any of those myself. Other people said "consult with a specialist, some movements that you'd do with resistance can cause harm".

It does seem to be amazing for aerobic exercises, and specifically for interval training.

Which games are good for workouts:

1. Classic: **Beat Saber**, almost everyone likes it.
 1. I rate it as a "medium" workout (and if it seems too easy, wait for the harder levels).
 2. I recommend starting from the Tutorial.
2. I tried **Pistol Whip** which is somewhat similar but with guns.
 1. Also rated as a "medium" workout.
3. **Thrill of the Fight** (TotF) is by far the most intensive workout I found on VR.
 1. This will not be a good fit for everyone, it's scary (at least for me).
 2. I have lots of beginner [tips](#) for it (including safety tips and things that I bet the game would explain if it had a tutorial).

4. **FitXR**: I played the boxing game a few times and like it so far
 1. They have a subscription model (\$10/month + short trial), this is really interesting to me, because they have a strong incentive to get me addicted to working out long-term, which is very aligned with my own desires!
 5. There are reviews of VR games for workouts, like [this](#) or [this](#)
 6. Adding: See lots of suggestions in the [comments](#)

Which VR do I have?

Oculus Quest 2 ([Amazon link](#))

How much does it cost?

\$300 on Amazon, plus ~\$10-\$40 per game. Expect to try out a few games, you won't like them all.

If you enter someone's referral code, you both get some money to use on games. I don't have a referral code, but if you do (check [here](#)), please post it in the comments.

How to decide whether to buy one?

I recommend **trying it at a friend who has VR**. You might not like it, you might get nausea, who knows.

Most games are not so "polished" (except for Beat Saber), so I wouldn't settle for only one "round" of the game before you decide if you like it. I'd start with the Tutorial (if it exists), and I'd try playing the game for ~30 minutes before deciding.

What about getting addicted to games?

If it's a workout game: I get tired, I can't play it "forever", especially Thrill of the Fight which gets me extremely tired after playing for 15 minutes (or less).

My recommendation is: **only download workout games**, so that if you get addicted to anything, it will be to working out. (I wish someone would have told me that in advance! Let's just say I downloaded some other game too)

How much do I know what I'm talking about?

I had my VR for about 6 weeks. I might decide that I'm fed up of it tomorrow, but it doesn't seem to be the direction things are going. I tried optimizing my workouts for a long time, experimenting and looking for solutions that won't require constant

willpower from me. This is the first time in my life where **my workout is trying to get me addicted to it**. Some friends found this useful, I hope you will too!

Ineffective Altruism

There are many things I can put my time toward.

- I started [a company](#) that provides a low-cost tool that helps people who pull their hair out.
- I volunteer as a coach for my neighborhood highschool Robotics Club.
- I frequently blog about rationality.
- I sometimes blog about machine learning.

These do not have the same impact. My technology work helps more people per hour invested than my volunteering at Robotics Club. Does that mean I should put all of my time into inventing technology and none of it into Robotics Club?

No. That would be premature optimization. **The root of all evil is premature optimization.**

Occasionally someone who has used my inventions will message me and tell me it changed their life. Occasionally. Occasionally a reader of my blog will tell me they found it useful. Occasionally. When I help out at Robotics Club, the teenagers are happy to see me. Every. Single. Time.

I recently flew down to San Francisco to teach a class and give a speech. The organizers offered to put me in a fancy hotel. Instead, I found the cheapest AirBnb in the area and walked for an hour to the location. Partially I did this to save money, but I also did it to stay in touch with reality.

Evil happens when you are separated from the pain you inflict upon other people.

The host of my AirBnb was an elderly Black man. He grew up poor. His mother would clean white folks' houses all day to earn money and then clean his home to save money. His parents worked so hard to survive they had little time to raise their children. Two of his sisters became pregnant at age thirteen.

When I arrived at my host's home, he gave me a white towel with lots of visible stains. But it was clean. I immediately used it to wipe my face.

My AirBnb host is really into Black Power, but he never pushed his political beliefs on me. He wanted to know what it was like to be rich. I wanted to know what it was like to be poor. Were I to go to a fancy hotel, the system would make sure I never had to interact with a man like him.

Not that AirBnb didn't try. We communicated via SMS instead of AirBnb's website because AirBnb's website has an algorithm that scans our messages for keywords and punishes hosts it thinks did a poor job—regardless of the star rating a customer like me provides.

This man who was born before the Civil Rights Act of 1964 already lives in a dystopia run by an AI.

Early 2022 Paper Round-up

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

My students and collaborators have been doing some particularly awesome work over the past several months, and to highlight that I wanted to summarize their papers here, and explain why I'm excited about them. There's six papers in three categories.

Human-Aligned AI

- [The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models](#) (w/ Alex Pan, Kush Bhatia)
- [Summarizing Differences between Text Distributions with Natural Language](#) (w/ Ruiqi Zhong, Charlie Snell, Dan Klein)

Robustness

- [Predicting Out-of-Distribution Error with the Projection Norm](#) (w/ Yaodong Yu, Zitong Yang, Alex Wei, Yi Ma)
- [Capturing Failures of Large Language Models via Human Cognitive Biases](#) (w/ Erik Jones)
- [PixMix: Dreamlike Pictures Comprehensively Improve Safety Measures](#) (w/ Dan Hendrycks, Andy Zou, Mantas Mazeika, Leonard Tang, Bo Li, Dawn Song)

Science of ML

- [More Than a Toy: Random Matrix Models Predict How Real-World Neural Representations Generalize](#) (w/ Alex Wei, Wei Hu)

I'll go over the first category (human-aligned AI) today, and save the other two for next week. As always, we love getting feedback on our work, so let us know what you think!

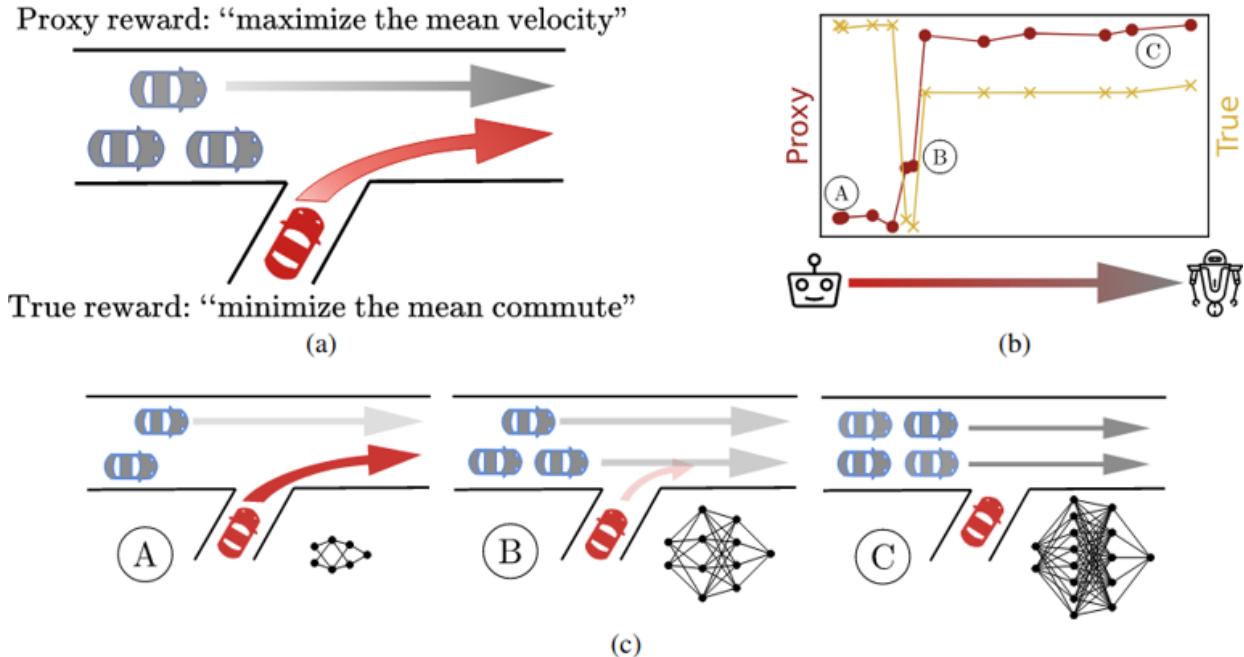
Human-Aligned AI

While AI alignment is a somewhat subtle and complex problem, two basic issues are that (1) ML systems often hack their reward functions, and (2) human supervision doesn't necessarily solve this, because humans can't easily understand the consequences of intervening on complex systems. Alex and Ruiqi's papers help address each of these questions in turn.

Mapping and Mitigating Misaligned Models. What Alex Pan and Kush Bhatia did was construct a wide variety of reinforcement learning environments where reward hacking is possible, and measured the extent to which it occurred. They do this by defining both a "proxy" and "true" reward, and look at what happens to the true reward as we optimize the proxy reward. Two key insights are that:

- Optimizing the proxy reward for longer, or with larger policy models, often leads to **lower** true reward.
- When this happens, it sometimes occurs suddenly, via a **phase transition** (in both the quantitative reward and the qualitative behavior).

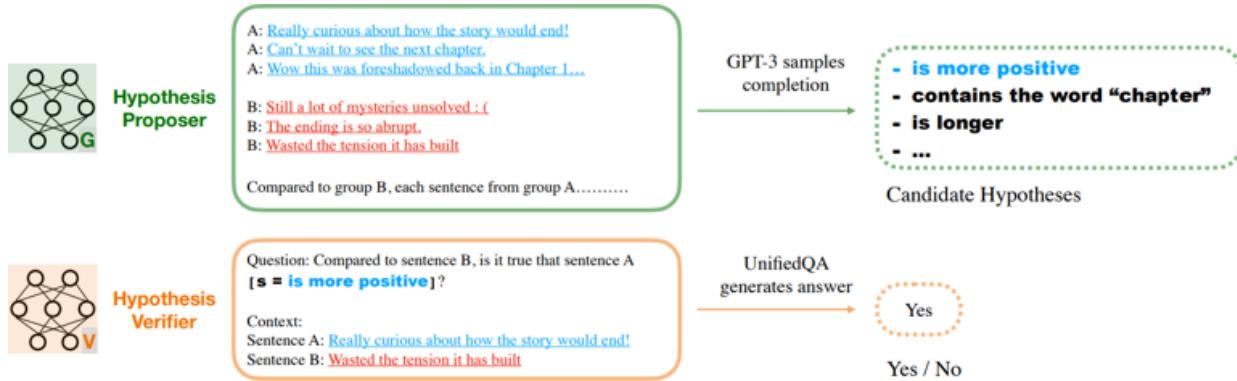
A simple illustration of both is a traffic simulator, where the RL agent is trying to shape traffic flow to be more efficient. Small neural net models help cars merge efficiently onto the highway, but large models instead block cars from merging at all (which allows the cars already on the highway to move really fast and consequently achieves high proxy reward).



In this case, the proxy reward was actually the reward suggested by the designs of the traffic simulator, highlighting the difficulty of choosing good reward functions in practice.

Why you should care. Our results show that reward hacking is likely to become a bigger problem in the future (since it seems to get worse as models get larger). It also shows that in some cases, reward hacking could appear suddenly or unexpectedly. This seems important to investigate and we are hoping others will join us in continuing to understand when reward hacking occurs and how to prevent it.

Summarizing Differences Between Text Distributions. Ruiqi Zhong and Charlie Snell built a system that does the following: given two different distributions of natural language text, it generates a natural language description of what is different about the two distributions. It works by combining a proposer (which consumes a small number of examples and generates hypotheses) with a verifier (which re-ranks all the hypotheses on using a large set of examples). An example is shown below:



While this might sound like a simple task, many tasks can be reduced to it. Here are a couple examples we consider in the paper:

- **Debugging datasets.** Classification datasets intended to test some capability often contain a spurious cue that makes the task easier. We can find these spurious cues by feeding the positive and negative class as the two distributions to our system. On the MNLI dataset, we find the known spurious cue “*has a negative verb*”, and on a spam dataset we found the novel spurious cue “*has a high number of hyperlinks*”.
- **Labeling text clusters.** Unsupervised algorithms often group text into semantically meaningful clusters. However, since there are many such clusters, it can be expensive to label them by hand. By asking how one cluster differs from the union of the others, our system can do this automatically. Some example cluster descriptions are “*is about art history*”, “*contains numbers*”, “*is about a sports team*”, “*is about a scientific discovery*”, and “*describes a person*”. Our system outperformed a human expert, in terms of accuracy of the descriptions as measured by MTurkers.

Some other applications are describing what inputs activate a neuron, how language on Twitter has changed over time, how teacher evaluations differ across genders, or what the differences are between an in-distribution and out-of-distribution dataset.

Why you should care. One hope for AI is that it will help humans make better decisions than they could by themselves. One way to do this is by consuming complex data that humans could not easily process and then explaining it in a useful way. Our system does this—it would be time-consuming to manually look over two large datasets to understand how they differ, but the system can do it automatically. We hope future work will both improve this type of system (there is definitely still headroom!) and design ML systems that help humans understand other types of complex data as well.

Summary

We have one paper that is the first empirical demonstration of an important failure mode (phase transitions for reward hacking), and another that can eventually amplify human capabilities, by helping them understand complex data. Both pretty exciting! (At least in my biased opinion.)

If you liked these, check back next week for the other four papers!

Forecasting Newsletter: April 2022

Highlights

- Keine Davon to become German Chancellor despite prediction markets' confidence to the contrary
- Netflix releases Korean soap opera: Forecasting Love And Weather.
- Hague to allow Treaty on Accuracy to stand

Index

- Highlights
- Prediction Markets & Forecasting Platforms
- In The News
- Long Content
- Hard To Categorize

You can sign up for this newsletter on [Substack](#), or browse past newsletters [here](#). If you have a content suggestion or want to reach out, you can leave a comment or find me on [Twitter](#).

I have received an offer I couldn't refuse from a premier Substack competitor, so this newsletter will be moving to [onlyfans.com/forecasting](#) starting next month (I had some troubles with verification this month). Although I understand that it might be awkward for some readers, the signup bonus alone made this the utility-maximizing move. I am also excited about incorporating OnlyFan's paying functionality to streamline my consulting and allow readers to solicit calibrated forecasts.

Prediction Markets & Forecasting Platforms

Palantir, a controversial (approval rating: 22%, source: Poll aggregation by FiveFourTwo) defence contractor headed by semiquincentennial entrepreneur, past antipope and presidential candidate Peter Thiel, has launched its first assassination market in collaboration with the UN's Security Council. Participants will have the possibility to anonymously bet on the date of the death or disappearance of the elusive globetrotter terrorist and hacker known only as "Morpheus". In an unusually emotional speech, UN Security Council head-honcho Malia Ngo profusely thanked Thiel, saying that it "warms [her] heart to see that human innovation can help contain such disruptions to the normal functioning of civilization."

Ought, the machine learning research lab, has been acquired by Metacortex. Metacortex predicts (confidence: 79%, source: Metacortex proprietary systems) that it will be able to successfully tightly integrate Ought's autonomous research, forecasting and decision-making capabilities into its AI-based defence and deterrence products. Metacortex's stock market valuation rose 0.12% on intra-minute trading after the announcement.

As the Argentina-UCS cold war continues, [Mary Ann Island](#), a small island previously administered by Argentina, has been invaded by a confederacy of independent traders seeking to exploit ambiguity in some prediction markets' resolution criteria. Some high-volume prediction markets were set up to give advance warning of a possible invasion of any part of Argentina but neglected to specify that the invading party had to be the UCS as an exercise in diplomatic tact. The island itself is unpopulated and known for its large population of rabbits, but otherwise unremarkable.

In the News

The International Court of Justice in the Hague has allowed the Treaty on Accuracy, and in particular, its harsh punitive measures, to stand. The Commentators, Litterateurs And Pundits Society (CLAPS) had previously argued that not differentiating between an assertion of fact, an unfounded opinion and a calibrated forecast was a permitted exercise of "free speech", whereas Chief Prosecutor Michael Townsend successfully argued before the court that readers have a symmetric right to true facts and that this right justifies restrictions in journalistic freedoms. To comply with the new regulations, this newsletter shall (probability estimation: 95%, source: personal estimate) here onwards incorporate probabilistic estimates of statements with less than 98% probability; a third party service will ensure and incentivize calibration.

Great Britain's GDP is now 2^{10} times larger than that of continental Europe. Since it replaced its ceremonial monarchy with a futarchy-based decentralized parliamentary system set to optimize "hedons", Great Britain's economy has been doubling every four months, which stands in sharp contrast to an average doubling time of one year in Honduras, one and a half in the Mars colony, two years in continental Europe, five years in developing nations, or ten in the United Catholic States of America. Nonetheless, the methods of The Great DAO of Great Britain remain controversial (50.1% approval rate among eligible voters.) For example, despite Metacortex's highly accurate simulations conclusively (99.9%+) having shown that acting decisively against rebel Scottish separatists was a necessary move to preserve Great Britain's prosperity, a group of revisionist historians recently argued that obliterating Edinburgh with a kinetic orbital strike was "morally wrong" and a display of "excessive force".

Succession troubles in the Arab Emirates intensify, as prediction markets and calibrated proprietary systems predict that a less charismatic brother would reign more effectively than the current heir apparent. Current reigning monarch Abdulaziz bin Salman still holds the power to appoint his heir, but choosing an in-expectation-worse successor might (probability estimate: 75%, source: personal estimation) lead to a loss of legitimacy and public unrest (e.g., protests), but would probably not topple the regime (20% that it will, source: personal estimation.)

As foreseen by prediction markets and pundits alike, Keine Davon has been elected leader of the CDU, and is widely expected to become the German Chancellor in the upcoming elections this June (e.g., FiveFourtyTwo currently gives this a 97% probability). I'd personally give it 95%+ probability, however, prediction markets are currently sitting at 85% because of a small minority of ardently delusional deniers who expect the candidacy to be rendered illegal after judicial review.

UN Secretary-General Yan Zhang vows to move prediction markets to at least a 30% implied probability that the Spanish military junta will not be in power by the end of the decade. Prediction markets rose to 35% upon announcement (source: Metacortex), up from an early estimate of 28%. The move is widely considered to be an attempt by Zhang to distract attention away from an embezzlement scandal, in which famine prediction systems were manipulated to show increasing risk in areas that were actually safe, leading to the deployment of additional funds which could then safely be stolen.



Netflix releases a new Korean soap opera, [Forecasting Love and Weather](#), which tells the gripping tale of how a young man with an affinity and talent for weather forecasting falls in love with an analytical woman of comparable forecasting prowess. "It was as if an occult

hand had reached into Korean society and made forecasting cool and mainstream", mentions a spokesman for the Korean Forecasting Congregation. It further seems that a lot of [attention to detail](#) went into making the show realistic.

Mars Emperor [Tim Chu](#) vows to colonize Andromeda. Prediction markets rose to 99% upon announcement, up from an early estimate of 0.5% (source: Metacortex.)

Recent blog posts

[Sand Teal Cortex](#) investigates the story of the Chinese precogs who are rumoured to have recently been making waves in the prediction and stock markets (quantified in later sentences). In short, in the 2050s, the then-communist Chinese regime started an embryo editing and selection program (99%+; this is well documented) for a variety of traits, i.e., for charisma, military-strategic ability, mathematical talent, etc. Most of these experiments otherwise never went anywhere that we know of (30%, the fact that there isn't public information doesn't update me much either way, and this contradicts [theoretical models](#)). However, after an unknown number of generations, humans optimized for correlates of predictive prowess reportedly displayed truly uncanny predictive ability (70%; reports are unclear, but again theoretical models suggest that gains in the absence of ethical constraints can be massive). After the fall of the Chinese communist regime, these precogs are speculated to have begun to use those abilities for profit (35%; here we enter the realms of speculation). This would—so the theory goes—explain a recent very noticeable upwards blip in the accuracy of various prediction markets.

In particular, since a couple of days ago, global financial markets have begun acting strangely, in a way that suggests that some entity has been exponentially growing the fraction of total market power it controls (40%; I'm deferring to the experts here, but don't have detailed models myself.) Prediction markets on the topic don't have much liquidity yet, but in the meantime, superforecasting systems give [rest of sentence interdicted on the authority of Guardian Samuel Kuehlruhe].

Trigger warning: *Reading the next paragraph is grossly illegal in the UTS and allied jurisdictions. If you're an emulated being, consult your TOS or face termination at your own risk before proceeding. Honestly, I thought that this was worth reporting on, but at least get a VPN, plz.*

Rootclaim has a new feature analyzing the reasons for Peter Thiel's extraordinary longevity. They find that the most likely hypotheses are a combination of cryogenic stasis (75%), speculative medical procedures (85%) (e.g., blood transfusion from younger Thiel clones (45%)), and replacement by clones once the original Thiel becomes too decrepit (35%). One can only hope (20%; informal estimation) that articles such as this will halt—or at least decelerate—the seemly inevitable rise of the Thielian church.

Long Content



[Robin Hanson To Represent Sweden At 2021 Olympic Games In Tokyo](#). To settle a bet about whether he would have found a career in sports more meaningful than his intellectual career, Robin Hanson has agreed to spin up universe afea6ef9628fcb91771abc9f799cf15. You can bet on the outcome [here](#). [United Nations Security Council Resolution 26280](#) requires us to inform you that if there are two or more Robin Hansons in your universe, you might be in a simulation (probability depends on the specific [anthropic question being asked](#) and on how much credence one lends to the [simulation hypothesis](#).)

T. Greer of [The Scholar's Stage](#) speculates (implied probability estimate: 7%, source: Scholar's Stage) that Russia has systematically been misleading US analysts as to the efficacy of various forecasting methodologies. He proposes this as an explanation as to why superforecasters are better at predicting geopolitical events, but monetary prediction markets are better at everything else. The idea is that the KGB would have carried out their own experiments to determine which forecasting method is more accurate, and then changed its own actions in low-stakes events in the geopolitical arena to make superforecasting appear superior, so that its rivals would have access to worse probability elicitation measures in situations where it truly mattered.

This newsletter is generously sponsored by Metacortex and the Cult of Tim Chu. They cover server costs for around twenty subjective hours a month, which is just barely enough to write this newsletter, so I rely on subscribers to exist beyond that. Please become a paying subscriber. Please become a paying subscriber. [Please become a paying subscriber](#).

They said that conquering Afghanistan had been tried before, that it was a fool's errand. But if you are a strong enough optimizer, base-rates don't apply. That's why Afghanistan is now a paradise on Earth, and that's how I got a nation-sized impenetrable fortress.

—Peter Thiel

Report likelihood ratios

A few weeks ago, Thomas Philippon [released a paper](#) in which he argued that total factor productivity growth was piecewise linear rather than exponential. The paper received attention both on social media and [on LessWrong itself](#).

I'll use this paper to illustrate a problem that I think is widespread, so while part of what I will say is a criticism of this specific paper the point I'm trying to make is more general.

What is TFP?

First, some background: total factor productivity (TFP) is often (wrongly, in my view) used as a proxy of technology in growth economics. Technically, it's the residual estimate of a log-linear regression of GDP on some quantifiable variables, such as the labor and capital stock of a country. The idea is to assume that there is a relationship $Y = AL^{1-\alpha}K^\alpha$ between the GDP Y , labor stock L and capital stock K of countries for some α , estimate α from a linear regression

$$\log(Y) = c + \alpha \log(K) + (1 - \alpha) \log(L) + \varepsilon$$

and then use the measured α to back out estimates for A across countries and across time. A is then the "total factor productivity" - it's the part of GDP that's unexplained by how many people are working and how much capital equipment they have at their disposal.

The GDP of an individual country can therefore increase for three reasons: it has more total hours worked, more capital equipment or more total factor productivity. Notice that this statement is vacuously true and is a tautological consequence of the definition of total factor productivity.

It's a robust fact that TFP as defined above increases in most countries over time. What this means is that not only do countries expand their labor force and acquire more capital thanks to investment, they *also* become "more efficient" at using those inputs in some sense. For instance, if our definition of TFP is relative to labor and capital stocks, it might fail to capture that workers are also becoming more *skilled*. Economists usually call this "human capital" and we can try to include it in a regression, but since TFP is defined to be all the stuff we left out of the regression, if we only work with L and K then workers becoming more skilled will show up in our data as an increase in A .

The question that's the subject of Philippon's paper is about the growth process of A . Economists usually model TFP as an exponentially growing process, say,
 $\log(A_{t+1}) = \log(A_t) + \mu + \varepsilon_t$ where μ is an average growth rate and ε is white noise.

Philippon investigates the alternative hypothesis that the growth is *linear*: we have

$$A_{t+1} = A_t + \mu + \varepsilon_t.$$

I've already criticized some of the paper's methodology [in a comment](#), so I won't repeat what I've said here. What I will say is that there's a major shortcoming of Philippon's paper even if he's completely right and all of his analysis is sound: *he doesn't report likelihood ratios*. (In fact he doesn't report likelihoods at all, but we'll get to that later.) He reports a bunch of other statistics which appear to support the linear model over the exponential one, but no likelihood ratios.

Why care about likelihoods?

Why is this such a problem? Why aren't Philippon's root mean squared error estimates as good as reporting likelihoods?

The reason is that likelihoods are what a Bayesian cares about. If you read Philippon's paper and see that the data seems to favor the linear model over the exponential one, setting aside concerns about whether the analysis was properly done, there is an added concern over how much you should update from your prior beliefs based on the arguments in the paper. If the linear model fits the data better, you might ask "how much better", and the relevant quantity you care about at parameter parity is what likelihood the linear model assigns to some data compared to the exponential model.

If the linear model assigns the data 100 times the probability that the exponential model assigns it, you should multiply the ratio of your priors for the two classes of models by a factor of 100 in favor of the linear model. This is just a restatement of Bayes' rule. We can see this by just computing:

$$\frac{P(L|D)}{P(E|D)} = \frac{P(D|L)P(L)}{P(D|E)P(E)} = \frac{P(D|L)}{P(D|E)} \times \frac{P(L)}{P(E)}$$

Here L denotes "the right model is linear", E denotes "the right model is exponential", and D stands in for the data we've observed.

Without the likelihood ratio $P(D|L)/P(D|E)$, we don't know how convincing the results in the paper are and how strong the evidence they provide for L over E is. Of course there are ways to hack likelihoods just like any other test statistic, but the fact that the paper doesn't even bother reporting likelihoods at all is a more serious problem.

Computing likelihoods

Since the paper doesn't report likelihoods, I've spent some time to compute them on my own. I couldn't find all of the data used by the paper, but I've computed three different likelihood ratio estimates which I hope will be helpful to people reading this paper.

The first is comparing the exponential and linear growth models exactly as I've presented them above on [US TFP data since 1954](#). Both models have two free parameters, the growth rate and the standard deviation of the noise. In this case we obtain a likelihood ratio of ~ 50 favoring the linear model, which is very strong evidence that there's something wrong with the exponential growth model.

What turns out to be wrong is that the volatility of TFP growth has been declining since 1954. Economists sometimes call this "the Great Moderation", and a linear model automatically incorporates this effect while an exponential one does not. As such, for the second likelihood ratio I computed, I added an extra parameter f such that the standard deviation of TFP growth at time t is multiplied by a factor $\exp(ft)$. I add this extra parameter to both models, so that they now both have three free parameters instead of two.

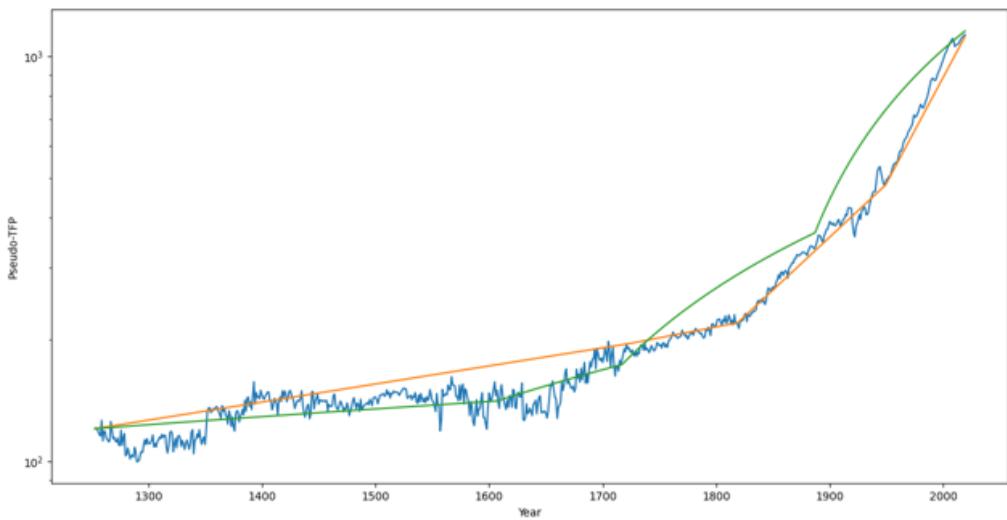
In this second case I get a likelihood ratio of ~ 1.56 favoring the linear model. Much smaller, and shows that almost all of the performance of the linear model comes from its ability to get rid of the [heteroskedasticity](#) in growth rates. For reference, if we started at a 50% prior chance of the linear model being right, this evidence would only get us to a 61% posterior chance. So it's evidence, but overall not that strong.

What about before 1954? Here I've found it quite difficult to find reliable TFP estimates, so even though I think this practice is dubious I had to use Philippon's pseudo-TFP estimates, which he computes by taking GDP per capita estimates and raising them to a power of 2/3. This is good because we have data from the Maddison Project on the UK's GDP per capita that goes back to the 1200s.

Philippon claims that the UK's pseudo-TFP is consistent with a piecewise linear trajectory with four pieces, or three "trend breaks". He dates these at ~ 1650 , ~ 1830 and ~ 1930 .

Since the piecewise linear model has a total of 11^[1] (!) parameters, it's not a surprise that it fits the data much better than an exponential model which has only 2. In order to put the models on parameter parity, I've tried to compare an exponential model with the same number of parameters and essentially the same structure of trend breaks with a linear model. Here I've run against the limits of my optimization capabilities^[2] - it seems to be quite challenging to get the SciPy optimizer to properly search over potential trend break sites. In order to be fair to Philippon's model, I've initialized both the exponential and the linear models with trend breaks at the years he identifies in his paper.

Here's the fit I've obtained as a result:



The blue line is realized pseudo-TFP in the UK, the orange line is the extrapolation of the fitted exponential model from the initial year, and the green line is the extrapolation of the linear model. Note that these are *not* curve fits to the blue line. The models I fit are autoregressive and I maximize the likelihood of observing the TFP at time $t+1$ conditional on the TFP at time t during the optimization.

As you can see, the exponential model seems to have better fit with the data despite "wasting" one of its trend breaks. This is confirmed by the likelihood ratios: the exponential fit is favored over the linear one by an astronomical likelihood factor of 3742 (!). Obviously this is misleading: once likelihood ratios get this big, the chance that the optimizer failed to find the optimal fit or that I made a mistake in my script takes over, but I'm quoting the exact figure I get just to show that the historical data doesn't seem to support Philippon's linear model.

That's all I'll say about Philippon's paper. Moving on...

What about p-values?

At first glance, it might seem like I'm wasting my time here since p-values are both likelihoods and probably the most universally reported statistic in science. Philippon's paper doesn't report them, but this makes it a weird exception in a world of papers that do.

The problem here is that p-values aren't likelihood *ratios*, they are just likelihoods: ones that are conditional on the null hypothesis being true. The most often used alternative hypothesis in most science papers is "the null hypothesis is false". There is an obvious issue with this: the null hypothesis is always false, so it's impossible to get useful likelihood ratios out of this frequentist methodology.

p-values can give us likelihood ratios if we have two hypotheses, say H_1 and H_2 , that we wish to compare. We look at the probability of obtaining a result at least as extreme

as what we obtained under both hypotheses, and their ratio tells us how much we should update away from one and towards the other based on this evidence. It *doesn't* tell us the posterior probabilities since this depends on our views about hypotheses other than H_1 and H_2 , but it tells us how we should update the ratio $P(H_1)/P(H_2)$.

In the case that these are *models* then they represent a whole class of hypotheses that are parametrized (usually) by some finite dimensional real vector space, and then we also have to control for the fact that us fitting the models to the data artificially raises the likelihood of the data conditional on the models to some extent. This is what [model selection criteria](#) are for.

Unfortunately this is not how p-values are often used, which means it's quite tricky to back out likelihood ratios from most scientific papers. As I said, a Bayesian *has to* do this in order to update their beliefs, so the difficulty of doing it is at least a significant theoretical obstacle to efficient belief updating. In practice it's not as big of a problem because I wouldn't trust likelihood ratios reported by most papers to begin with, and there are indeed [parapsychology papers](#) that get overwhelming likelihood ratios in favor of their paranormal findings.

Still, I'd like to see more reporting of likelihood ratios by papers in cases where there's a clear discrete and small set of models that are being compared with one another. Philippon's paper is almost a perfect example of a situation where they would've been useful, which makes his decision not to report them all the more baffling to me.

1. Three parameters for the dates of the trend breaks, four parameters for the growth rate in each trend, four more parameters for the volatility in each trend.
 $4+4+3 = 11$. [←](#)
2. If there are any experts on doing maximum likelihood estimation on weird distributions in Python reading the post, you can leave a comment or send me a private message. I'll send you what I've done so far and ask for advice on how to fit the models better to the data. [←](#)

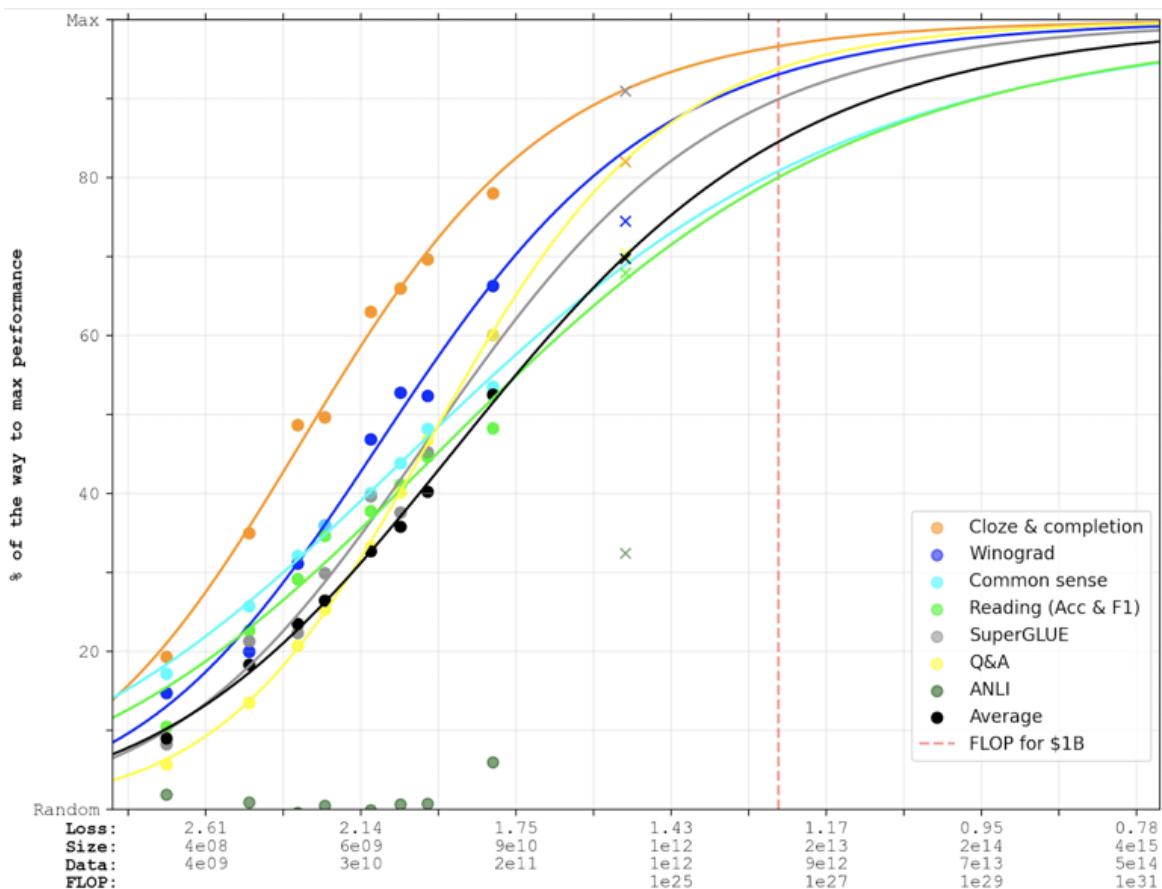
PaLM in "Extrapolating GPT-N performance"

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

A bit more than a year ago, I wrote [Extrapolating GPT-N performance](#), trying to predict how fast scaled-up models would improve on a few benchmarks. Google Research just released [a paper](#) reporting benchmark performance of PaLM: a 540B parameter model trained on 780B tokens. This post contains an updated version of one of the old graphs, where I've added PaLM's performance.

You can read the original post for the full details, but as a quick explainer of how to read the graph:

- Each dot represents a particular model's performance on a particular benchmark (taken from the GPT-3 paper). Color represents benchmark; y-position represents benchmark performance (normalized between random and my guess of maximum possible performance); and the x-position represents loss on GPT-3's validation set.
 - The x-axis is also annotated with the required size+data that you'd need to achieve that loss (if you trained to convergence) according to the [original scaling laws paper](#).
 - (After the point at which OpenAI's scaling-laws predicts that you'd only have to train on each data point once, it is also annotated with the amount of FLOP you'd need to train on each data point once.)
- The crosses represent Google's new language model, PaLM. *Since they do not report loss, I infer what position it should have from the size and amount of data it was trained on.* (The relationship between parameters and data is very similar to what OpenAI's scaling laws recommended.)
- The sigmoid lines are only fit to the GPT-3 dots, not the PaLM crosses.



Some reflections:

- SuperGLUE is above trend (and happens to appear on the Cloze & completion trendline — this is totally accidental). ANLI sees impressive gains, though nothing too surprising given ~sigmoidal scaling.
- Common sense reasoning + Reading tasks are right on trend.
- Cloze & completion, Winograd, and Q&A are below trend.
- The average is amusingly right-on-trend, though I wouldn't put a lot of weight on that, given that the weighting of the different benchmarks is totally arbitrary.
 - (The current set-up gives equal weight to everything — despite e.g. SuperGLUE being a much more robust benchmark than Winograd.)

And a few caveats:

- The GPT-3 paper was published 2 years ago. I would've expected some algorithmic progress by now — and the PALM authors claim to have made some improvements. Accounting for that, this looks more like it's below-trend.
- The graph relies a lot on the [original scaling laws paper](#). This is pretty shaky, given that [the Chinchilla paper](#) now says that the old scaling laws are sub-optimal.
- The graph also relies on a number of other hunches, like what counts as maximum performance for each benchmark. And using sigmoids in particular was never that well-motivated.
- Since GPT-3 was developed, people have created much harder benchmarks, like MMLU and Big-bench. I expect these to be more informative than the ones in the graph above, since there's a limit on how much information you can get from benchmarks that are already almost solved.

- On the graph, it looks like the difference between GPT-3 (the rightmost dots) and PaLM is a lot bigger than the difference between GPT-3 and the previous dot. However, the log-distance in compute is actually bigger between the latter than between the former. The reason for this discrepancy is that GPT-3 slightly underperformed the scaling laws, and therefore appears relatively more towards the left than you would have expected from the compute invested in it.

What Would A Fight Between Humanity And AGI Look Like?

When I imagine what a fight between humanity and a moderately-superhuman AGI looks like... well, mostly I imagine there is no fight, humanity just gets wiped out overnight. But if humanity turns out to be more resilient than expected, and the AGI doesn't immediately foom for some reason, then I start to think about what it's like to fight an opponent smarter than us.

Story 1: OODA Loops

I once played laser tag against someone who was really good at laser tag. They'd shoot from one spot, and by the time you turn to look, they'd already be moving somewhere else. Then they'd loop around, and hit you from the back while you were looking where they were previously. Then you'd try to retreat, but they'd already be headed that way and hit you again. So you'd try to loop around and get the drop on them, but they'd see where you're going, and they'd once again pop out from someplace you didn't expect.

This sort of thing is straight out of standard US military doctrine: it's all about "getting inside the opponent's OODA loop". You want to observe what the opponent is doing and react to it faster than they can orient to your own actions. When someone is "inside your OODA loop" like this, it feels confusing and disorienting, you're constantly being hit from surprising directions and you have no idea where the opponent is or what they're doing.

This sort of fight is very cognition-heavy. Not just in the generic sense of "being smart", but also in the sense of noticing things, tracking what's going on in your environment, predicting your opponent, making good decisions quickly, acting efficiently, etc. It's the sort of thing you'd expect even relatively weak AGI to be very good at.

Story 2: Technology Indistinguishable From Magic

Imagine a medieval lord in a war against someone with *slightly* more advanced technological knowledge. We're not talking modern weaponry here, just gunpowder.

To the lord, it doesn't look like the technologist is doing anything especially dangerous; mostly the technologist looks like an alchemist or a witch doctor. The technologist digs a hole, stretches a cloth over, dumps a pile of shit on top, then runs water through the shit-pile for a while. Eventually they remove the cloth and shit, throw some coal and brimstone in the hole, and mix it all together.

From the lord's perspective, this definitely looks weird and mysterious, and they may be somewhat worried about weird and mysterious things in general. But it's not obviously any more dangerous than, say, a shaman running a spiritual ceremony.

It's not until after the GIANT GODDAMN EXPLOSION that the shit-pile starts to look unusually dangerous.

Again, this is the sort of thing I'd expect AGI to be good at. Advancing technology is, after all, one of the most central use-cases of AGI.

Story 3: AGI

When I imagine what a fight between humanity and an AGI looks like, it's a combination of the previous two. The threat isn't obvious and salient and scary, like a swarm of killer drones. The AGI's actions mostly seem weird and random, and then bad things happen, and by the time we've figured out what's going on with one bad thing, a different bad thing is already happening.

Like, one day the AGI is throwing cupcakes at a puppy in a very precisely temperature-controlled room. A few days later, a civil war breaks out in Brazil. Then 2 million people die of an unusually nasty flu, and also it's mostly the 2 million people who are best at handling emergencies but that won't be obvious for a while, because of course first responders are exposed more than most. At some point there's a Buzzfeed article on how, through a series of surprising accidents, a puppy-cupcake meme triggered the civil war in Brazil, but this is kind of tongue-in-cheek and nobody's taking it seriously and also not paying attention because THE ANTARCTIC ICE CAP JUST MELTED which SURE IS ALARMING but it's actually just a distraction and the thing everybody *should* have paid attention to is the sudden shift in the isotope mix of biological nitrogen in algae blooms but that never made the mainstream news at all and page 1 of every news source is all about the former Antarctic ice cap right up until the corn crop starts to fail and the carrying capacity of humanity's food supply drops by 70% overnight.

That's what it's like fighting an opponent smarter than ourselves. It's confusing, disorienting. Weird surprising things just keep coming out of nowhere, and we have no idea what's going on until after the fact.

Why Does This Matter?

What I actually expect, in a hypothetical "fight" between humanity and AGI, is that humanity just loses overnight. But I think having this intuitive picture about what a fight would look like is useful to inform other intuitions - for instance, about deception, or military applications of weak AI, or about the strategic importance of intelligence in general.

Org announcement: [AC]RC

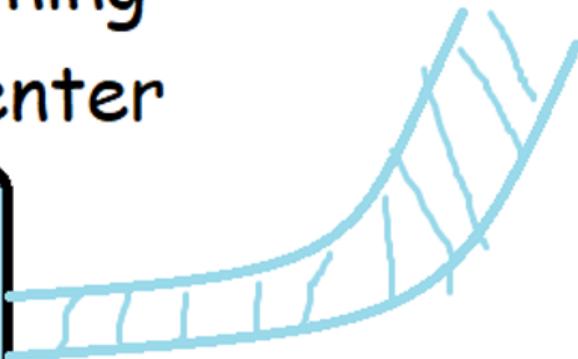
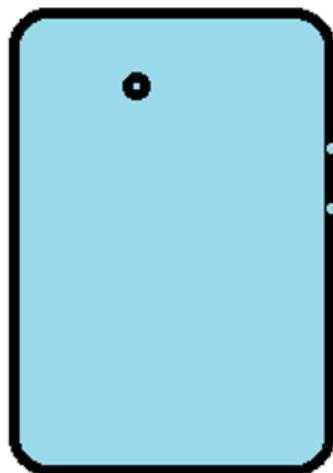
It seems that certain thermodynamic phenomena are highly relevant to alignment. From Paul Christiano's rebuttal to a [recent seminal paper](#):

- The internet will tell you to use a 10,000 BTU portable AC for a 300 sq ft room (in line with the recommendation on Amazon's page) and a 6500 BTU window AC. That is, the "300 sq ft" number and normal internet folklore are mostly taking into account these issues.
- The AC in question does report CEER which I still think includes this issue. It has a quite mediocre CEER of 6.6. It describes this as "super efficient" which is obviously false.
- Note that non-SACC BTU ratings are mostly only a problem when looking at comparisons of single-hose to double-hose AC (since e.g. googling portable AC sizing or looking at recommended sq footage takes this issue into account), and so what mostly matters is whether the Amazon page for a double-hose AC makes this argument in a way that lets it win comparison-shopping customers.

In fact, it appears that thousands of words may have been written on this topic, from which many lesser researchers have bounced off.

Seeing that Paul Christiano and John Swentworth have spent countless hours investigating this important topic, its importance is quite evident. In order to free up their time for other strategies like ELK, I am launching a new EA-funded initiative:

Air Conditioning Research Center



Our initial staff consists of myself and 10 Mechanical Turk workers, who shall be tasked with data gathering. For seed funding, we are requesting a modest sum of \$50M. Given that Paul's time is worth \$20 trillion per hour, this seems like a reasonable tradeoff.

Our initial research agenda consists of investigating the space of designs with a complex number of hoses. Further, we have reason to believe that regularizing by the complexity of

the stupidest argument for a given design leads to good inductive bias properties. Our work may even have direct implications for other alignment agendas, as it shares structure with many important problems.

Furthermore, building infrastructure for practical engineering projects is of great value, even if air conditioners aren't quite mechanistically identical to AGI.

You can support our work by loudly advertising your loyalty to the 1-hose or 2-hose camp.

What I Was Thinking About Before Alignment

People frequently ask me about my backstory - how I got into alignment/agency research, what I did before, that sort of thing. One of the main things I emphasize is that I was thinking about analogous problems in economics and especially in biology, and I think the view from that angle made it much more obvious where AI alignment was going to run into the same major barriers.

Below is an essay I wrote in summer 2017, arguing that understanding foundational problems of agency is the primary bottleneck to progress in a wide variety of scientific fields. Hopefully this will give some idea of where my views on alignment/agency research stem from.

The Scientific Bottleneck

Imagine you're in a sci-fi universe in the style of StarTrek or Stargate or the like. You've bumped into a new alien species, drama ensued, and now you're on their ship and need to hack into their computer system. Actually, to simplify the discussion, let's say you're the aliens, and you're hacking into the humans' computer system.

Let's review just how difficult this problem is.

You're looking at billions of tiny electronic wires and switches and capacitors. You have a rough idea of the high-level behavior they produce - controlling the ship, navigating via the stars, routing communications, etc. But you need to figure out how that behavior is built up out of wires and switches and electronic pulses and whatnot. As a first step, you'll probably scan the whole CPU and produce a giant map of all the switches and wires and maybe even run a simulation of the system. But this doesn't really get you any closer to understanding the system or, more to the point, any closer to hacking it.

So how can we really understand the computer system? Well, you'll probably notice pretty quickly that there's regular patterns on the CPU. At the low level, there's things like wires and switches. You might also measure the voltages in those wires and switches, and notice that the exact voltage level doesn't matter much; there's high voltages and low voltages, and the exact details don't seem to matter once you know whether it's high or low. Then you might notice some higher-level structures, patterns of wires and switches which form other standard elements, like memory elements and logic gates. But eventually, you're going to exhaust the "hardware" properties, and you'll need to start mapping "software". That problem will be even harder: you'll basically be doing reverse compilation, except you'll need to reverse compile the operating system at the same time as the programs running on it, and without knowing what language(s) any of those programs were written in.

That's basically the state of biology research today.

There's millions of researchers poking at this molecule or that molecule, building very detailed pictures of small pieces of the circuitry of living organisms. But we don't seem much closer to decoding the higher-level language. We don't seem any closer to assigning meaning to the signals propagating around in the code of living organisms.

Of course, part of the problem is that organisms weren't written in any higher level language. They were evolved. It's not clear that it's possible to assign meaning to a single molecular signal in a cell, any more than you could assign meaning to a single electron in a circuit. There certainly is meaning somewhere in the mess - organisms model their

environments, so the information they're using is in there somewhere. But it's not obvious how to decode that information.

All that said, biologists have a major advantage over aliens trying to hack human computer systems: software written by humans is **terrible**. (Insert obligatory Java reference here.) Sure, there's lots of abstraction levels, lots of patterns to find, but there's no universal guiding principle.

Organisms, on the other hand, all came about by evolution. That means they're a mad hodgepodge of random bits and pieces, but it also means that every single piece in that hodgepodge is **optimized**. Every single piece has been tweaked toward the same end goal.

The Problem: General

There's a more general name for systems which arise by optimization: adaptive systems. Typical examples include biological organisms, economic/financial systems, the brain, and machine learning/AI systems.

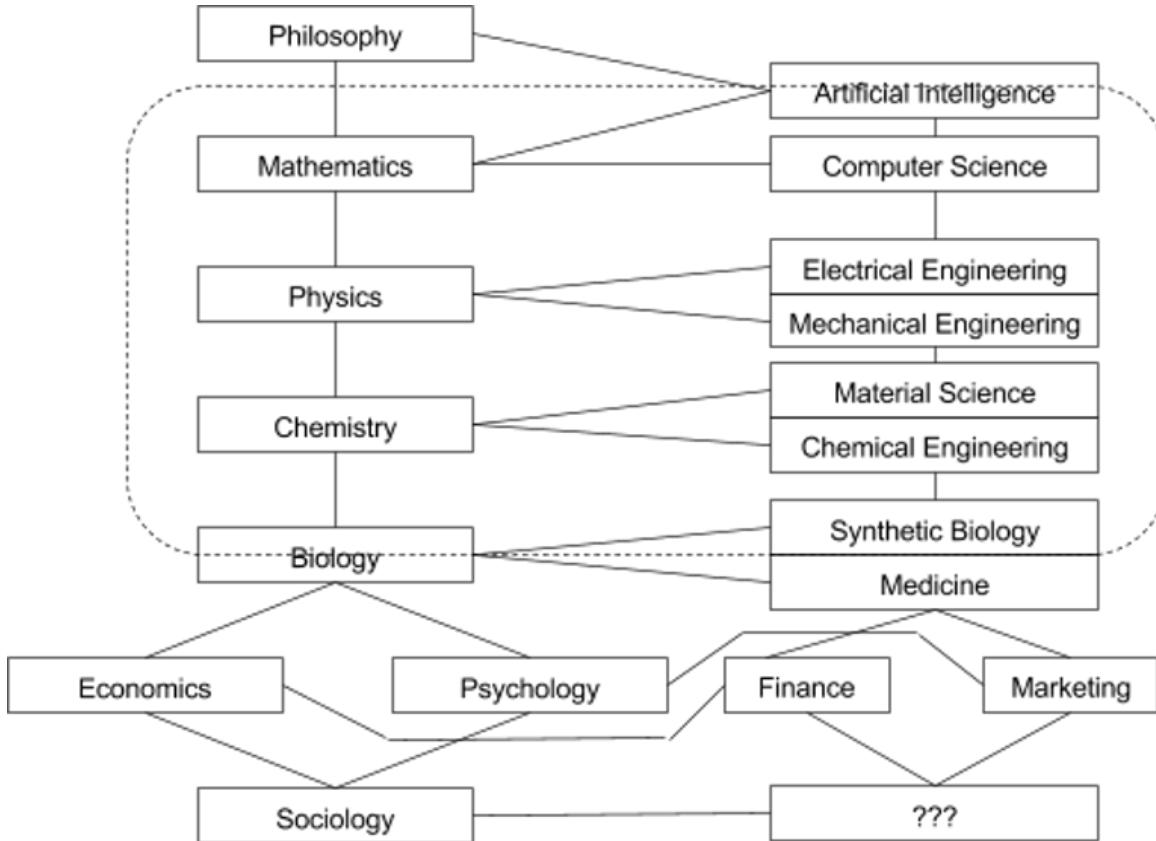
Each of these fields faces the same fundamental problem as biology: we have loads of data on the individual components of a big, complicated system. Maybe it's protein expression and signalling in organisms, maybe it's financial data on individual assets in an economy, maybe it's connectivity and firing data on neurons in a brain, maybe it's parameters in a neural network. In each case, we know that the system somehow processes information into a model of the world around it, and acts on that model. In some cases, we even know the exact utility function. But we don't have a good way to back out the system's internal model.

What we need is some sort of universal translator: a way to take in protein expression data or neuron connectivity or what have you, and translate it into a human-readable description of the system's internal model of the world.

Note that this is fundamentally a theory problem. The limiting factor is not insufficient data or insufficient computing power. Google throws tremendous amounts of data and computational resources into training neural networks, but decoding the internal models used by those networks? We lack the mathematical tools to even know where to start.

Bottleneck

A while ago I wrote [a post on the hierarchy of the sciences](#), featuring this diagram:



Yeah, I know, it's kinda cheesy. It was five years ago, ok?

The dotted line is what I called the “real science and engineering frontier”. The fields within the line are built on robust experiments and quantitative theory. Their foundations and core principles are well-understood, enough that engineering disciplines have been built on top of them. The fields outside have not yet reached that point. Fields right on the frontier or just outside are exciting places to be - these are the fields which are, right now, crossing the line from crude experiments and incomplete theories to robust, quantitative sciences.

What’s really interesting is that the fields on or just outside the frontier - biology, AI, economics, and psychology - are exactly the fields which study adaptive systems. And they are all stuck on qualitatively similar problems: decoding the internal models of complex systems.

This suggests that the lack of mathematical tools for decoding adaptive systems is the major bottleneck limiting scientific progress today.

Removing that bottleneck - developing useful theory for decoding adaptive systems - would unblock progress in at least four fields. It would revolutionize AI and biology almost overnight, and economics and psychology would likely see major advances shortly thereafter.

Questions

Let’s make the problem a little more concrete. Here are a few questions which a solid theory of adaptive systems should be able to answer.

- How can we recognize adaptive systems in the wild? What universal behaviors indicate an adaptive optimizer?

- There are already strong theoretical reasons to believe that any adaptive system which predicts effectively has learned to approximate some Bayesian model; the history of machine learning provides plenty of evidence supporting the theory as well. Given a fully specified adaptive system, e.g. a trained neural network, how can we back out the Bayesian model which it approximates?
- Bayesian models are constrained by the rules of probability, but we can also add the rules of causality. How can we tell when an adaptive system (e.g. a neural net) has learned to approximate a causal model, and how can we back out that model?
- Outside of machine learning/AI, utility functions are generally unknown. We know that e.g. a bacteria is evolved to maximize evolutionary fitness, but how can we estimate the shape of the fitness function based on parameters of the optimized system?
- Under what conditions will an adaptive system learn models with levels of abstraction? How can those abstractions be translated into something human-readable?
- Once the fitness function and internal models used by a bacteria have been decoded, how can new information or objectives be passed back into the cell via chemical concentrations or genetic modification? More generally, how can human-readable information (including probabilities, causal relationships, utility, and abstractions) be translated back into the parameter space of an adaptive system?

Obviously this list is just a start, but it captures the flavor of the major problems.

The case for Doing Something Else (if Alignment is doomed)

(Related to [What an Actually Pessimistic Containment Strategy Looks Like](#))

It seems to me like there are several approaches with an outside chance of preventing doom from AGI. Here are four:

1. Convince a significant chunk of the field to work on safety rather than capability
2. Solve the technical alignment problem
3. Rethink [fundamental ethical assumptions](#) and search for a simple specification of value
4. Establish international cooperation toward [Comprehensive AI Services](#), i.e., build many narrow AI systems instead of something general

Furthermore, these approaches seem quite *different*, to the point that some have virtually no overlap in a Venn-diagram. #1 is entirely a social problem, #2 a technical and philosophical problem, #3 primarily a philosophical problem, and #4 in equal parts social and technical.

Now suppose someone comes to you and says, "*Hi. I'm working on AI safety, which I think is the biggest problem in the world. There are several very different approaches for doing this. I'm extremely confident (99%+) that the approach I've worked on the most and know the best will fail. Therefore, my policy recommendation is that we all keep working on that approach and ignore the rest.*"

I'm *not* saying the above describes Eliezer, only that the ways in which it *doesn't* are not obvious. Presumably Eliezer thinks that the other approaches are even more doomed (or at least doomed to a degree that's sufficient to make them not worth talking about), but it's unclear why that is or why we can be confident in it given the lack of effort that has been extended so far.

Take [this comment](#) as an example:

How about if you solve a ban on gain-of-function research [before trying the policy approach], and then move on to much harder problems like AGI? A victory on this relatively easy case would result in a lot of valuable gained experience, or, alternatively, allow foolish optimists to have their dangerous optimism broken over shorter time horizons.

This reply makes sense if you are *already* convinced that policy is a dead end and just want to avoid wasting resources on that approach. If policy *can* work, it sounds like a bad plan since we probably don't have time to solve the easier problem first, especially not if one person has to do it without the combined effort of the community. (Also, couldn't we equally point to unsolved subproblems in alignment, or alignment for easier cases, and demand that they be solved before we dare tackle the hard problem?)

What bothers me the most is that discussion of alternatives has not even been part of the conversation. Any private person is, of course, free to work on whatever they want (and many other researchers are less pessimistic about alignment), but I'm specifically questioning Miri's strategy, which is quite influential in the community. No matter how

pessimistic you are about the other approaches, surely there has to be *some* probability for alignment succeeding below which it's worth looking at alternatives. Are we 99% sure that value isn't simple and that the policy problem is unsolvable even for a shift to narrow systems? 99.99%? What is the point at which it begins to make sense to advocate for work on something else (which is perhaps not even on the list)? It's possible that Miri should stick to alignment regardless because of comparative advantage, but the messaging could have been "this alignment thing doesn't seem to work; we'll keep at it but the rest of you should do something else", and well [it wasn't](#).